II SUPPLY OF HEALTH CARE

5 THE LABOR MARKET FOR PHYSICIANS

A 59-year-old woman on her walk to work unexpectedly falls to the sidewalk and is unable to get up. A passerby calls for an ambulance and she is rushed to the hospital. An emergency-room doctor examines her and finds her conscious but groggy. Her speech is slurred, and she is unable to move her right arm or leg. The doctor concludes that she has almost certainly suffered a stroke.

Now the doctor must make a vital decision on her behalf. Broadly speaking, there are two distinct types of strokes: clotting and bleeding. One possibility is that a blood clot in the patient has lodged in the wrong place. This is called an ischemic stroke. Another possibility is that an artery in her brain, for whatever reason, has ruptured and she is bleeding internally. This is called a hemorrhagic stroke. Both cut off blood supply to the brain.

The diagnostic problem that the doctor faces is that these two possibilities require dramatically different treatments. If it is a clot that has caused the stroke, the correct course of treatment is to give her a drug that dissolves the clot. If the treatment is applied in time, the damage from the stroke may be minimal and reversible. Unfortunately, if the cause is a ruptured blood vessel, this course of treatment is counterproductive and may even kill her.

The doctor must decide what to do quickly with incomplete information. The patient entrusts the doctor with her life – she is in no position to help decide, and even if she were healthy she has no experience distinguishing between an ischemic and hemorrhagic stroke.

From the patient's point of view, it would be best if the doctor focuses only on the medicine and leaves out of his decision-making all matters extraneous to her health and well-being. The patient hopes the doctor does not consider his own financial well-being and that he leaves aside any prejudices he might have. Most of all, the patient hopes that her doctor knows what he is doing. In her groggy state, she cannot tell how capable her doctor is, so she has to trust that regulatory agencies have barred unqualified physicians from practicing. But such regulations may come with the cost of more expensive health care and shortages in doctor supply.

The economic literature on the physician labor market emphasizes tradeoffs that are implicit throughout this story. How should doctors be trained to minimize harm to patients? How should the patient be compensated for mistakes that doctors make, and what consequences will that have for how doctors practice? If doctors are allowed to set the required standard for what it means to be a high-quality doctor (and who else would know enough to do so?), does this mean they will earn monopoly rents? Our goal in this chapter is to make clear the tradeoffs inherent in regulating the market for physicians.

5.1 The training of physicians

There is a consensus among nations that, before they can practice, physicians must receive an education in both basic sciences and clinical training. And in all countries, this training requires several years to complete. But nations differ in how long the training takes, and how physicians must demonstrate their competency.

Medical school

In most European countries, aspiring physicians enter medical school directly out of high school. Conversely, almost all medical schools in the US and Canada require a bachelor's degree for admittance, so students there must first attend an undergraduate institution. They can pursue a major or concentration in any field they choose, but to apply for medical school they also must complete a pre-medical curriculum of biology, physics, chemistry, mathematics, and English.

Entry into medical school is a selective and competitive process. Only about one in two people who apply to US medical school each year are accepted to any schools, though this rate fluctuates and has been as low as one in three in years past. Entry into the most selective schools is even more difficult. For instance, the University of California, San Francisco, received 6,767 applications in 2011 for only 149 slots. In the UK too, the admissions process is quite selective. In 2004, there were over twice as many applicants as spaces at the UK medical schools (Powis et al. 2007).

Given the difference in the ages in which students enter medical school, the length of medical school predictably differs across countries as well. In the US, it typically lasts four years; in the UK, five years; and in France, students need up to six or seven years to fully complete their studies. In all cases, the first portion of medical school focuses mostly on classroom work: students study topics like anatomy, physiology, pharmacology, pathology, and biochemistry.

During the second half of medical school, focus shifts from classroom work to clinical and patient management skills. Students serve as junior members on teams taking care of patients in teaching hospitals and go through monthly rotations on different teams in specialty areas such as internal medicine, surgery, pediatrics, and gynecology. These rotations help students learn about the different branches of medicine and prepare students to select a specialty.

Though the subject matter is similar across medical schools, the tuition cost shouldered by students varies tremendously across countries. Four years of medical training in the US can cost around $140,000 at public schools and $225,000 at private ones (Morrison 2005). Meanwhile, both the German and French governments heavily subsidize medical school so that students pay only €200 to €500 a year (Segouin et al. 2007; Chenot 2009). The cost of medical schools has an impact on who attends medical school and the makeup of the labor market of physicians. We discuss this effect in Section 5.2.

Residency

Medical school is only the beginning of the process of learning to be a doctor. Graduates fresh out of medical school are not qualified to take care of patients on their own; they have spent years in the classroom learning anatomy and pharmacology, but

relatively little time caring for patients. One of the main jobs of doctors is helping patients make life-and-death decisions with incomplete information and often under great time pressure. No classroom experience by itself can equip someone to perform such tasks.

Becoming a doctor thus requires an extensive, hands-on apprenticeship, called a residency in the US. In the final year of medical school, students pick a specialty within medicine – such as surgery, pediatrics, or internal medicine – and then apply to residency programs that train physicians in that specialty. Residency is an intense and arduous time for young physicians. Work weeks typically extend up to 80 hours and sometimes more, and a single shift can last 36 hours or longer. Most of the learning in residency takes place under the supervision of veteran doctors called attending physicians who help trainees perform surgeries, make clinical decisions, and manage patients.

Upon completion of their first year of residency, called an internship, doctors can earn their license to practice medicine, which also gives them the right to prescribe drugs. Though legally permitted to practice at this point, few doctors leave residency after their internship, because insurers are reluctant to reimburse incompletely trained doctors for their services.

After residency, which typically takes at least three years and can take much longer depending on the specialty, some doctors continue their medical training in a subspecialty (Figure 5.1). For example, in order to practice as a cardiologist, a doctor must complete a three-year internal medicine residency and then a cardiology fellowship, which may take an additional five years.

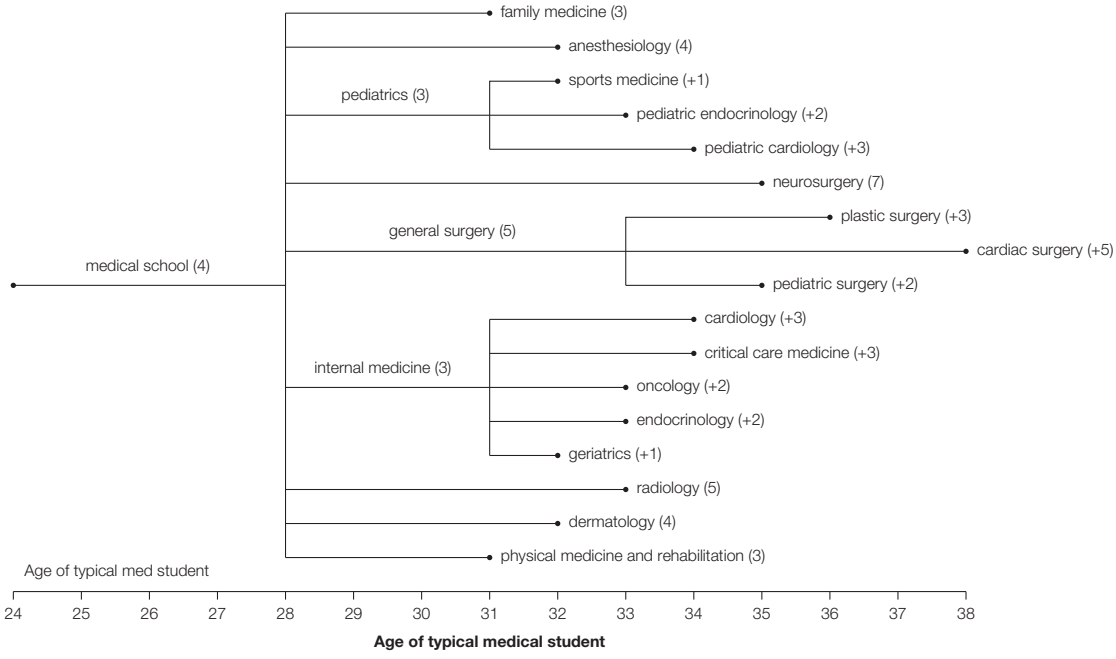


Figure 5.1. *Time requirements in years for various residency and fellowship programs in the US.*

Source: Data from GME Directory 2010–2011.

Resident inexperience and the July effect

On-the-job training is an inherent part of the residency system. The unofficial motto that guides much of medical training is "See one, do one, teach one." Although this process is largely unobserved by patients, even the greenest residents routinely participate in medical procedures and surgeries, including highly complex ones. As Atul Gawande underscores in *Complications* (2002), any surgeon who has mastered a procedure originally had to perform that same procedure for the very first time – likely on a patient who had no awareness of what was happening.

Such on-the-job experience is clearly indispensable to training competent and experienced doctors, but it is inevitable that inexperienced residents will make some mistakes. American medical professionals have long spoken of a "July effect," a reference to the month of the year when experienced residents finish their terms and fresh medical school graduates take their place (Young et al. 2011). Because experience levels decline abruptly in American teaching hospitals in July, some fear that the quality of medical care available at these hospitals plummets during the first few weeks of the new term. In the UK, where residency programs transition later in the summer, the same phenomenon is called the "August killing season."

If the July effect is real, Americans should be very careful handling fireworks on Independence Day (July 4). Credit: © Carlos Santa Maria – Fotolia.com.

Empirically, there is substantial evidence that the July effect is real. For instance, US counties with teaching hospitals experience a 10% spike in fatal medication errors in July, while counties without teaching hospitals do not (Phillips and Barker 2010). Some studies fail to find a significant effect on mortality (Huckman and Barro 2005; Bakaeen et al. 2009), but a comprehensive review of the literature suggests otherwise. In a review of 39 recent studies of resident changeover, Young et al. (2011) find substantial evidence that patient mortality increased and hospital efficiency suffered in the periods immediately following a changeover.

Physician work-hours

Physicians are among the longest-working professionals; a doctor's work-"day" can last more than 24 hours. There are patients to treat, operations to conduct, and innumerable forms to complete. Even a single surgery can last longer than a typical 9-to-5 work-day: a surgeon performing a pancreaticoduodenectomy, for example, may be in the operating room for ten straight hours.[1]

So it is unsurprising that physicians frequently clock more than 60 work-hours in a week (Dorsey et al. 2003). The time spent in the hospital may be even greater for residents. Historically, a resident on call overnight once every three nights would routinely work up to 30 hours consecutively and 96 hours or more in a single week (Steinbrook 2002).

1 A pancreaticoduodenectomy, commonly known as a Whipple procedure, is performed to remove part of the pancreas and part of the small intestine.

The fatigue of working long hours may impair a physician's cognitive abilities, which in turn may have adverse effects on patient health. Taffinder et al. (1998) and Eastridge et al. (2003) compare the performance of well-rested surgeons and sleep-deprived ones on a virtual-reality laparoscopic surgery simulator. Taffinder et al. find that surgeons with no sleep needed 14% more time and committed 20% more errors than surgeons who operated on the simulator after a full night of sleep. Likewise, Eastridge et al. find a significant difference in the number of errors.

A converse hypothesis argues that longer work-hours for physicians and residents actually has benefits for patient health in the long run. If doctors work continuously for long periods, a hospitalized patient may remain with the same doctor for her entire stay. This not only improves the patient experience, but more importantly, requires fewer hand-offs between different physicians, thereby minimizing the chance that crucial information is mishandled (Arora et al. 2005).

In 2003, the Accreditation Council for Graduate Medical Education (ACGME) implemented limits on the number of hours that residents in the US are permitted to work. Doctors-in-training were not to work in excess of 24 consecutive hours and no more than 80 hours in a single week.[2]

The ACGME ruling created a natural experiment to estimate the impact on patient outcomes of limiting resident work-hours, as only teaching hospitals with residency programs were affected by the new limits. If health outcomes changed in teaching hospitals but not in non-teaching ones, researchers could identify the difference as the result of the new ACGME limits. However, studies found little to no difference in mortality outcomes in the two years following the reform (Volpp et al. 2007; Prasad et al. 2009). Shetty and Bhattacharya (2007) do find a small decrease in mortality risk for medical patients, but no significant change for surgical patients. These results suggest that constrained resident hours may improve outcomes for patients in teaching hospitals, but the effects are not large. However, the studies consider only patient deaths – they do not track possible changes in the number of non-fatal errors.

One explanation for the lack of a measurable effect is that the ACGME's policy may not have been completely effective at limiting work-hours. One survey found that 83.6% of survey respondents violated the regulation at least once during the year after policy implementation (Landrigan et al. 2006).

Landrigan et al. (2004) conducted a randomized experiment at the Brigham and Woman's intensive care unit (ICU) at Harvard to study the effect of work-hour limitations on physician errors. The experiment randomly assigned two teams, one to a traditional schedule and one to a shorter work week. They closely monitored each team to ensure that neither deviated from its assigned schedule. Those in the traditional schedule averaged 77 to 81 hours a week and up to 34 consecutive hours. The alternative schedule limited shifts to a maximum of 16 hours and averaged 60 to 63 work-hours a week. A physician observer accompanied each intern to chronicle (and intercept) medical errors.

2 One of the authors, who was in medical school before these limits were established, once toiled 112 hours in an obstetrics ward in a single week. He was exhausted by the end but three new mothers were grateful for his efforts – or at least the efforts of the more experienced doctors who were there to correct his mistakes.

They found that interns on the traditional schedule committed 35.9% more serious medical errors than the interns on the limited schedule. That difference included 20.8% more medication errors and 5.6 times as many diagnostic errors. Patient outcomes, however, did not differ significantly between the two groups, because senior physicians intercepted most serious errors.

This highlights a policy tradeoff. Residency is designed to be a safe place for new doctors to gain experience and make errors without harming patients. An unintended consequence of work-hour limits may be that residents get less supervised practice before the end of their residency. Though no study has examined this tradeoff explicitly, it may be the case that work-hour restrictions reduce errors by residents at the expense of increasing errors by recent graduates of residency programs.

Other nations have also grappled with the tradeoffs in imposing restrictions on resident work-hours (Woodrow et al. 2006). In Canada, resident unions have negotiated for and won work-hour legislation in each province. While the exact ceiling varies from province to province, they roughly match the 80-hour limit in the US (Romanchuk 2004).

Physicians in Europe, meanwhile, are subject to the European Union's Working Time Directive (WTD), which establishes a maximum 48-hour work week. This edict, aimed at ensuring quality of life for European workers, has raised worries about shortages in health services and reduced opportunity of training for doctors (Sheldon 2004; Maxwell et al. 2010). However, an independent review commissioned by England's National Health Service in 2010 argues that even under the WTD proper training can be delivered as long as effective supervision is given (Temple 2010).

5.2 Physician wages

The content and length of physician training is similar across nations. But physician wages vary tremendously across countries. In the US, physician salaries are high compared with salaries in other professions. But becoming a physician requires years of expensive training, both in medical school and during residency. As Adam Smith (1776) observed in *Wealth of Nations*, "wages vary with the cost of learning the business." In order to see if this fact fully explains the high salaries of working physicians, we must consider income over the entire lifespan.

Returns to medical training

Consider a student who is thinking of becoming a doctor but is also considering another career, such as professional surfing. If he decides to become a surfer, he can start earning money right away. On the other hand, becoming a licensed physician requires four years of medical school as well as additional years of residency training. In the US, four years of medical school can cost between $140,000 and $225,000, though in most European countries it is free or heavily subsidized. Then physicians go through a period of low-paying residency before becoming fully licensed to practice. Only at this late date does the physician begin earning a high salary. Figure 5.2 portrays the two income paths for professional surfers and licensed physicians in the US and in western Europe.

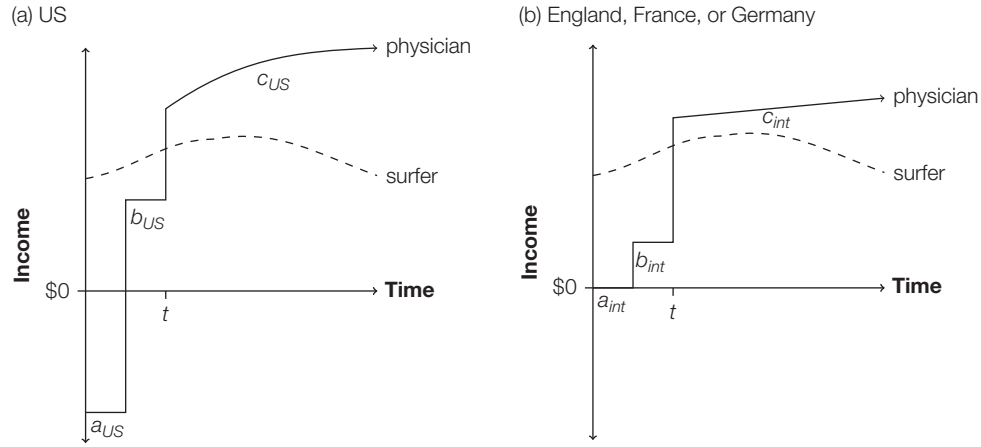


Figure 5.2. *The surfer earns a moderate income over his entire career. (a) In the US, an aspiring physician earns negative income while in medical school (interval a_{US}) and then relatively low income during residency (interval b_{US}). After that though, he makes a very high income (interval c_{US}). (b) In countries where medical school is entirely or heavily subsidized, physicians-in-training sacrifice less income early on (interval a_{int}) but also tend to earn less after graduation.*

As long as the college graduate enjoys ripping waves about as much as he enjoys seeing patients – that is, so long as the only relevant difference between the professions is the income stream – he should pick the career that maximizes the **net present value** (NPV) of his income. In this context, NPV captures how valuable a lifetime stream of income is to an individual today. It depends on the total amount of income, the individual's discount rate, and the degree of backloading in the income stream. Recall that the discount factor δ is a number between 0 and 1 that indicates how much a person values utility in future periods and in essence how much patience that person has. A higher δ indicates that the person values future utility nearly as much as present utility, implying greater patience.

Definition 5.1

Net present value (NPV): the discounted sum of all future earnings from $t = 0$ onward. The discount factor δ is between 0 and 1 and is a measure of how much less the individual values future income than present income:

$$\text{NPV} = \sum_{t=0}^{T} \delta^t I(t) \tag{5.1}$$

The discount *factor* δ can be expressed also as a function of a discount *rate* r:

$$\delta = \frac{1}{1+r} \tag{5.2}$$

Expressed in this way, high values of r indicate impatience. It is convenient to express δ this way, because the discount rate r presents a direct comparison with market interest rates. Suppose someone has a discount factor of $\delta = 0.9$. Then applying equation (5.2)

yields $r = 0.11$. This person would rather spend his money than save it unless the bank offers an interest rate greater than 11%.

"I actually just couldn't stand the sight of blood."
Credit: © AZP Worldwide – Fotolia.com.

The returns to a surfing career are frontloaded and the returns to a career in medicine are backloaded, so the more impatient college graduate will pursue surfing while the patient one pursues medicine.

College graduates with high values of discount rate r, and therefore low discount factors δ, are impatient and value the quick returns of a surfing career more than the delayed returns from a career in medicine. Graduates with low values of r are patient and are more likely to pursue medical careers. Thus, as discount rate r moves from 0 to 1, the person is more and more impatient, and therefore more likely to go into surfing. This also means that somewhere between 0 and 1, there is a specific discount rate r^* such that the person is indifferent between surfing and medicine. At r^*, the person values the financial returns from both careers equally. This r^* is known as the **internal rate of return** (IRR) of going into medicine versus surfing.

Definition 5.2

Internal rate of return (IRR): the discount rate r^* of an investment that would imply a net present value of 0 for that investment versus the best alternative. If $I_p(t)$ is the income in period t from the investment, and $I_s(t)$ is the income in period t from the alternative, then the internal rate of return r^* satisfies the following equation:

$$\sum_{t=0}^{T} \frac{I_p(t)}{(1+r^*)^t} - \sum_{t=0}^{T} \frac{I_s(t)}{(1+r^*)^t} = 0 \tag{5.3}$$

If the internal rate of return is *higher* than an individual's discount rate, he should make the investment – the future returns are worth the wait. If the internal rate of return is *lower* than an individual's discount rate, he should not make the investment – the future returns are not worth the wait (see Figure 5.3).

There is a long history of economists interested in calculating the IRR of a career in medicine. Nobel laureate Milton Friedman wrote his dissertation in the 1930s on measuring the IRR of becoming a physician versus a lawyer in the US (Friedman and Kuznets 1945). Burstein and Cromwell (1985) find that internal rates of return for entry into medicine or dentistry in the US were over 10% in the 1970s, even when accounting for differences in work-hours (Table 5.1). Other recent estimates using data from the 1990s indicate that the IRR has only grown over time (Weeks and Wallace 2002).

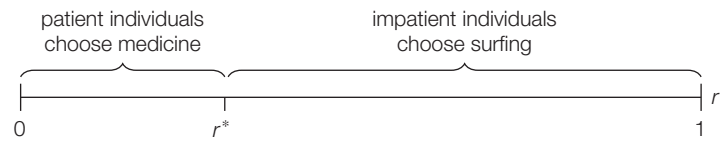


Figure 5.3. *Career choice depends on discount rate. If the internal rate of return r^* increases, more people will be inclined to become physicians.*

Table 5.1. *Estimated IRR for various professional careers versus a typical college-degree-requiring job in the US, 1970–1980.*

Year	All physicians	General practitioners	Dentists	Lawyers*
1970	11.8%	12.1%	16.1%	7.0%
1971	11.6%	13.2%	—	6.6%
1972	10.7%	12.2%	14.4%	5.7%
1973	10.8%	12.5%	—	6.7%
1974	12.0%	14.5%	14.9%	7.1%
1975	11.6%	12.3%	—	7.1%
1976	10.5%	12.4%	15.8%	7.1%
1977	10.2%	13.3%	—	6.8%
1978	11.0%	13.0%	16.3%	6.8%
1979	11.6%	14.5%	—	7.2%
1980	12.1%	14.2%	—	7.2%

Internal rates of return were calculated in reference to the earnings of the population of college graduates, and adjusted for differences in work-hours.

*The internal rates of return reported for lawyers are *not* adjusted for differences in work-hours.

Source: Burstein and Cromwell (1985). With permission from Elsevier.

In fact, in the US the internal rates of return for a professional career are substantially higher than the market interest rate. By itself, this means that these professional careers are good investments. This presents an economic puzzle: How can the rates of return for entering these careers remain so high? Presumably people would be attracted to these high returns, so more and more people would decide to go to medical school. A glut of new doctors would compete salaries down to a point where the IRR of becoming a physician is no longer higher than the market interest rate.

The fact that the IRR has instead remained high suggests that there are obstacles to entry into medicine. First, it may be the case that few people have an aptitude for medical practice, so many people attracted to medicine would not be able to find jobs as physicians even with medical training. Alternatively, the supply of physicians may be constrained by the number of slots in medical schools and restrictions in the immigration of physicians. We discuss these possible obstacles further in Section 5.3.

Returns to specialization

Almost all doctors are well paid, but doctors within certain specialty fields – like surgery, cardiology, and radiology – command even higher salaries (Table 5.2). Entering these specialties requires newly minted physicians to complete longer residency periods. So doctors graduating from medical school and selecting an area of specialty face a dilemma similar to that of the college graduate discussed earlier: Are they willing to defer income in the short term (that is, complete a longer residency) in order to enjoy higher income later?

It turns out that pursuing a highly paid specialty field is almost always worthwhile, even considering the longer low-paid residency. Nicholson (2002) estimates that the internal

Table 5.2. *Survey data on average hourly wage and work-hours in various medical specialties in the US, 2004–2005.*

Specialty	Mean hourly wage	Mean work-hours/week
Neurologic surgery	$132	58
Immunology	$112	49
Orthopedic surgery	$108	61
Dermatology	$103	45
Gastroenterology	$93	57
General surgery	$86	61
Obstetrics and gynecology	$83	57
Psychiatry	$72	45
Geriatric medicine	$57	53
Internal medicine and pediatrics	$50	57

Source: Data from Leigh et al. (2010).

rates of return for physicians entering into radiology, orthopedic surgery, general surgery, obstetrics/gynecology, and anesthesiology in 1998 exceeded 25%. He finds that entering these specialties would still be advantageous even if their residency positions did not pay *at all*.

Can the huge returns to high-paying medical specialties possibly be a competitive labor market outcome? It is actually possible for high salary differentials to persist in competitive equilibrium if . . .

- . . . *specialists work more*. If specialists work consistently more than non-specialists, then their extra wages are compensation for longer hours and the burden of working late nights or early mornings.
- . . . *specialists train longer*. If specialists have to spend many more years in residency than non-specialists, then higher wages are necessary to induce them to take those extra years to train.
- . . . *specialists retire earlier*. If specialists tend to retire earlier, either because their longer schedules wear them out or because their physical dexterity (a key attribute for surgeons) diminishes rapidly with age, then annual salary differentials are not the best way to compare specialists and non-specialists. The NPV or IRR comparisons are more appropriate, because they account for lifetime income.
- . . . *specialists are more highly skilled*. If specialists have rare abilities that make them particularly well suited to their demanding jobs, then they should enjoy a wage premium even in a competitive labor market.

Bhattacharya (2005) examines data on young physicians and uses an IRR calculation to find that differences in work-hours, training requirements, career length, and ability (as measured by test scores) can only explain half of the salary premiums for medical specialists. Surprisingly, differences in ability at the end of medical school do not explain any of the differences in returns between high-income and low-income specialties. Since these four explanations do not account for all of the difference between specialist and

non-specialist salaries, presumably the remaining differences must be due to barriers to entry to becoming a specialist.

5.3 Barriers to entry

Early in American history, the marketplace for physician services was largely unregulated and doctors were in plentiful supply. Kessel (1958) depicts a nineteenth-century medical marketplace that was unfettered and competitive:

> With very few exceptions, anyone who wanted to practice [medicine] was free to hang a shingle outside and declare himself available. Medical schools were easy to start, were easy to get into, and provided . . . a varied menu of medical training that covered the complete quality spectrum.

In the present day, the market for physicians is much more tightly regulated, with various restrictions on entering medical school and practicing medicine. The bulk of evidence suggests that these barriers to entry allow physicians to earn **monopoly rents**, extra wages above the competitive level that accrue to producers in a market where supply is artificially constrained. This section details those barriers and discusses the costs and benefits of policies that restrict physician supply.

Definition 5.3

Monopoly rents: extra wages above the competitive level that accrue to producers in a market where supply is artificially constrained.

The rise of the AMA

The conditions described by Kessel (1958) did not persist. In the nineteenth century, a movement to regulate medical practice began to take shape. Early attempts included an 1827 meeting of doctors in Northampton, Massachusetts, to agree upon certain basic standards for medical education and an 1835 effort by the faculty at the Medical College of Georgia to create an organization to oversee medical practice. In 1847, the American Medical Association (AMA) was founded and immediately took up the cause of ensuring that all practicing doctors had a "suitable preliminary education" and a uniform set of "elevated requirements" for the MD degree.

At first, the AMA focused its resources on an extensive licensure campaign. It persuaded states to require anyone who wanted to practice as a physician to obtain a license to do so. States mandated that physicians complete an examination or a certain amount of training to become licensed. While successful in convincing many states to license the practice of medicine, the AMA was not without detractors. Kessel recounts a speech by a Massachusetts state legislator who acknowledged the virtue of medical cures but compared doctors and medical professionals to "a powerful trade union [demanding] legislation against the competition of the 'scabs.' "

The AMA turned its focus to closing low-quality medical schools in the early twentieth century. The release of an influential 1910 report by Abraham Flexner – which corroborated a 1906 AMA finding that nearly half of the 160 medical schools then in

existence were of unacceptably low quality – was a turning point in American medical history. The Flexner report convinced state legislatures and medical examining boards to appoint the AMA as certifier of all MD-granting institutions. The number of medical schools and medical students began declining precipitously thereafter, and the number of medical students did not return to its 1910 level for over thirty years. Only in the twenty-first century has the number of US medical schools rebounded to the 1910 level.

Present-day barriers in the US physician labor market

In the decades since the Flexner report, the AMA, together with the American Association of Medical Colleges (AAMC), has consolidated control over the training process for US physicians. Together they run the Liaison Committee on Medical Education (LCME), which manages the pipeline of new physicians into the US labor market. These organizations decide which medical schools deserve accreditation, impose strict caps on medical school class sizes, and limit residency program enrollment. The caps on medical school class size are strict enough that less than half of all applicants were admitted to US medical schools in each year from 2006 to 2010 (Vassev and Geraci 2010). There may be a pool of qualified individuals who would like to become doctors but do not have the opportunity to even begin the training process.

For physicians to practice on their own, they must obtain a license. In the US, states grant licenses to physicians who have successfully completed medical school, the first year of residency, and a national licensing examination. This is a significant constraint because physicians who attempt to practice without a license can be thrown in jail.

International medical graduates (IMGs) from overseas medical schools face additional barriers to entering the American physician labor market (Educational Commission for Foreign Medical Graduates 2012). Those without the legal ability to work in the US must vie for a limited number of H1-B visas for highly skilled immigrants. They must pass the same licensing tests required of American medical graduates and complete a US residency (even if they have been practicing successfully overseas for years). Non-native English speakers must also pass a clinical skills assessment testing their ability to interact with English-speaking patients. Meeting these requirements is an onerous process and a logistical challenge. While most tests are available online, the clinical skills assessment is only offered in five US cities and requires a fee of $1,350 as of 2012 (applicants are responsible for their own travel expenses). Over 40% of IMGs who apply for licenses to practice medicine in America ultimately do not succeed.

Some have called for nurses and physician assistants to play a larger role in providing health care in order to alleviate physician shortages or bring down prices, especially in rural areas where doctors are particularly scarce (Pohl et al. 2011). But the growing number of nurses who provide services traditionally performed by physicians also face barriers. Most states allow nurse practitioners to provide certain kinds of care such as managing diabetic patients, suturing minor wounds, or prescribing physical therapy. There are limits, though: nurses are not allowed to prescribe medication and physician assistants cannot conduct major surgeries.

Insurers, including Medicare and Medicaid, have policies that make it harder for nurses to compete. Medicare and other insurers routinely pay nurses less than licensed physicians for providing the same service, and many states require nurses to be supervised or associated with a physician in order to get payment. These policies prevent nurses

from opening their own practices to compete directly with doctors, and put them at a competitive disadvantage.

Practitioners of alternative medicine, like chiropractors, homeopaths, and acupuncturists, also face barriers to providing services that compete with traditional, allopathic physician services (Anderson et al. 2000). All fifty states require chiropractors to be licensed, and many states also require acupuncturists to be licensed, further constraining the supply of health care providers.

The implicit tradeoff of barriers to entry

Barriers to entry in the labor market for physicians result in monopoly rents. Friedman and Kuznets (1945), Leffler (1978), Svorny (1987), and Anderson et al. (2000) all find evidence that various barriers to entry increase US physician salaries. Outside the US, the same story holds. Immigrant licensure policies create rents for physicians in Israel (Kugler and Sauer 2005) and a lottery for medical school admission in the Netherlands generates large rents for physicians there (Ketel 2011).

These rents, which are estimated to approach 25% of total physician compensation in some cases, are costly because they raise prices for anyone who seeks to purchase physician labor. In theory, a perfectly competitive labor market would dissipate these rents as more and more health care providers flood into the market, lowering prices for all consumers.

But training and licensure requirements for physicians have existed since ancient times (Leffler 1978). Why do they endure? Policies restricting entry into the physician labor market are defended by the AMA and others on the grounds that these policies assure high physician quality. The barriers are designed to ensure that anyone who completes medical school and becomes licensed to practice medicine is a qualified doctor. Unlike in the days before the AMA, dangerously unqualified people cannot offer medical services to unwitting patients.

In the typical market, consumers are given the prerogative to make determinations about supplier quality and make purchases accordingly. Consumers who demand high quality, pay for it with higher prices, while others may opt for lower-quality alternatives if the price is right. The market for physicians may be atypical though, because physicians offer a suite of complex services. If it is costly for consumers to distinguish good physicians from bad on their own, it may be worthwhile for consumers to defer to regulatory bodies like the AMA to decide which physicians should be allowed to practice.

Is it worth paying more for physician services (in the form of monopoly rents) if it means that consumers do not have to spend time or money investigating which physicians are qualified? This is the tradeoff a society weighs when it considers a licensure regime for physicians. If the search costs to consumers in an unregulated market exceed monopoly rents in a regulated market, then barriers to entry may improve welfare after all.

5.4 Physician agency

By stepping into a doctor's office, the patient is anointing the physician as an agent for his health. One major role of physicians is to advise patients about their health in settings where patients have very little technical knowledge about medicine. Physicians thus have a professional responsibility to serve as good stewards for their patients since it is difficult

for patients to assess whether the doctor is doing the right thing. The patient hopes that the doctor will conduct himself "only for the good of [the] patients, keeping . . . far from all intentional ill-doing," in keeping with the Hippocratic Oath. Doctors who follow this dictum are good agents for their patients' health.

Physicians face incentives that may cause them to deviate from perfect agency. For financial, legal, or personal reasons, the physician might overprescribe procedures, underprescribe treatment, or not treat certain patients altogether. The next sections discuss why physicians may deviate from perfect agency in theory and practice. Chapter 15 explores these themes in the context of health policy.

Physician-induced demand

One of the primary things patients pay doctors for is information. Doctors are trained to interpret subtle signs and make treatment recommendations on that basis. Unless the patient is savvy about matters of health – what certain diagnoses mean and what the proper treatment should be – he is likely to be swayed by the doctor's recommendations. In a sense, the uninformed patient is at the mercy of his doctor.

This information asymmetry between doctor and patient creates an opening for doctors to prescribe more services than patients want or need. Doctors may not serve as perfect agents for their patients. Patients want good advice and treatment that maintains or restores their health. Doctors want this, but they may also design treatment with an eye toward other goals such as their own financial well-being. If so, a doctor may induce a patient to demand more than the patient would want if he were better informed. This phenomenon is called **physician-induced demand** (PID).

Definition 5.4

Physician-induced demand (PID): extra demand for medical goods and services induced by the advice of a physician who takes into account goals other than the patient's objectives, such as the physician's own financial gain (McGuire 2000).

Financial gain might motivate physicians to induce demand, but there are also costs to inducement (otherwise, doctors would induce without restraint or remorse). Evans (1974) posits that physicians suffer psychic costs such as guilt as a result of misleading patients away from their self-interest. Ethical norms and a particularly strong sense of professionalism among doctors tend to discourage inducement. Stano (1987a) draws an analogy between inducements and advertising. Like advertising, inducement increases demand but also consumes resources. So even the profit-maximizing doctor does not seek infinite inducement.

Another impediment to inducement is competition. When physicians have market power, they have strong incentives to induce, because they earn economic profits on each patient treated. In a competitive market, profits are zero, so there is nothing to gain from inducement to offset the psychic costs (Stano 1987b).

The empirical evidence suggests that physicians do indeed change their practices in response to financial incentives. Hickson et al. (1987) randomly assigned a group of pediatric residents to be either paid by a fixed salary or paid according to the number

of services rendered. If these physicians were good agents for patient health, then we would expect the treatment patterns of the two groups to be statistically indistinguishable, because payment structure has nothing to do with patient health outcomes. But the evidence shows that fee-for-service physicians not only scheduled more visits per patient than salaried physicians did, they also scheduled more visits than were recommended by the American Academy of Pediatrics.

A separate observational study found that orthopedists and neurologists who owned their own MRI machines ordered more MRI tests than did physicians who had to refer patients to outside firms for MRI scans. Under the PID theory, the financial gain from billing for more diagnostic tests lured doctors into inducing more demand for MRI scans (Baker 2010). Mitchell (2008) tells a similar story about physician ownership of specialty treatment clinics. Physicians who recently became owners of back and spine clinics were more likely to recommend surgery than they were before owning their own clinics. Yip (1998) finds that thoracic surgeons compensated for a round of cuts to Medicare fees in 1990 by performing more surgeries.

The PID phenomenon poses a challenge for insurers who manage physician incentives through reimbursement policies. Setting high reimbursements could lead doctors to overprescribe certain procedures. But lowering reimbursements might cause physicians to switch patients to other, more lucrative services. The quandary of PID motivated Medicare actuaries in the late 1980s to lower surgical fees by an additional 6.5% to account for the anticipated volume inducements by physicians. The lack of evidence that volume actually increased for surgeries in subsequent years could suggest either that any effect of PID is small or that physicians opted to induce demand for more lucrative services whose fees had not been reduced (McGuire 2000).

Defensive medicine

PID occurs when physicians overprescribe treatment for personal financial gain, but there are other reasons why doctors might deviate from optimal medical practice. One theory proposes that doctors order superfluous tests or procedures to reduce the risk of medical malpractice lawsuits. This may manifest in extra diagnostic tests, low-value treatments, and even unnecessary invasive procedures. Doctors may also reject or refer away high-risk patients (or litigious ones) to reduce malpractice risk. This is known as **defensive medicine**.

Definition 5.5

Defensive medicine: deviations from optimal medical practice in order to reduce the risk of conflict with patients, especially in the form of malpractice lawsuits.

Fear of malpractice suits is not unfounded: medical errors do occur and invite liability lawsuits. Jena et al. (2011) find that between 1991 and 2005, 7.4% of all US physicians covered by a large liability insurer faced at least one malpractice claim in any given year. From these annual percentages, they extrapolate that 75% of physicians in low-risk specialties and 99% in high-risk specialties will have suffered at least one malpractice claim by the end of their careers. The average payout on successful claims was $274,887, though

78% of claims did not result in payouts. Even unsuccessful malpractice claims can still be injurious to physicians, as they still have the power to damage reputations and can require time and resources to defeat.

Physicians widely admit to practicing defensive medicine in this liability environment. Among high-risk practitioners in Pennsylvania, 93% of doctors surveyed reported practicing defensive medicine, 59% conducted more diagnostic tests than they thought were medically necessary, and 39% avoided caring for high-risk patients in order to reduce their exposure to liability (Studdert et al. 2005). Doctors practice defensive medicine in Europe too, even though malpractice litigation risk is lower there than in the US. Surveys of Dutch family physicians found that 17% of referrals and 27% of tests were ordered for defensive reasons, largely due to a desire to avoid personal conflicts with patients (Veldhuis 1994).

Survey data is self-reported, so the percentages might understate the degree of defensive medicine if physicians are not consciously defensive, or they might be inflated if physicians misidentify optimal medical behavior as defensive. Other researchers have relied on changing legal climates as natural experiments to measure the extent of defensive medicine. Throughout the 1980s, many states instituted caps on the damages patients could recover through malpractice suits as part of a broader movement towards tort reform. Some states adopted these reforms in the early days of the movement, while others adopted them later or not at all.

Several researchers have used this wave of tort reforms in the 1980s to study defensive medicine. Kessler and McClellan (1996) find that lower liability pressure as a result of the reforms led to a 5–9% decline in medical expenditures for patients with serious heart diseases. The lower expenditures did not produce significant effects on health outcomes. Helland and Showalter (2006), studying the same reforms, estimate that a 10% decrease in expected liability costs leads doctors to increase their work-hours by 2.85%. Apparently, capping liability did reduce the incentives of doctors to practice defensive medicine and even led some doctors to increase their caseload.

Mello et al. (2010) estimate that the financial costs as a result of the medical liability system in the US reached $55.6 billion annually in 2008 dollars or 2.4% of US health care spending. Their estimate includes the costs of actual payouts from malpractice claims, attorney fees, and higher medical costs due to defensive medicine, but omits social costs like the damage to reputation from liability suits, which are more difficult to quantify.

Another way that doctors respond to liability risk is through the purchase of liability insurance. Liability insurance, like health insurance, smooths consumption by charging a premium to doctors and reimbursing those doctors who face malpractice claims.[3] If liability insurance were full and could completely insure physicians from all the costs of malpractice claims, then there would be little incentive for doctors to practice defensive medicine.

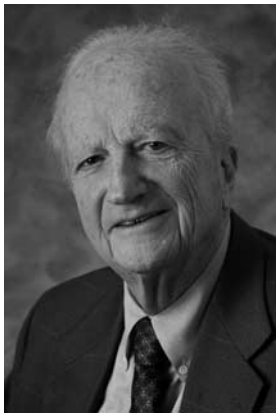
Full liability insurance might induce another perverse incentive though. Rather than being overly cautious about care, fully insured doctors might be less diligent in protecting against errors, because the risk of malpractice now is shouldered entirely by their insurer. If so, full liability insurance might lead to more medical errors, more malpractice claims, and more malpractice payments. This is why liability insurers are reluctant to offer full malpractice insurance packages.[4]

---

3 See Chapter 7 for a more complete discussion.

4 See Chapter 11 for more about this tradeoff.

5.5 Racial discrimination by physicians



1992 Nobel laureate Gary Becker pioneered the application of economic principles to topics in sociology, including the study of racial discrimination. He is famous for what is known as the Rotten Kid Theorem. *Credit: Gary Becker.*

The difficulty patients face in assessing physician behavior opens the door for another pernicious consequence of physician agency – racial discrimination. If physicians harbor racial animus or act on unfounded stereotypes, minority patients may lose trust in their doctors' advice or receive worse care. Researchers have uncovered differential treatment rates by race in heart attack care, organ transplantation, and many other areas of medicine (Werner et al. 2005; Liu et al. 2011). Some have even proposed that racial discrimination by physicians is responsible for the health disparity between different racial groups noted in Chapter 4. However, racial differences in treatment may be an optimal response driven by biological or behavior differences, at least in some cases.

Types of discrimination

Discrimination by physicians might take a few different forms. Maybe some doctors prefer not to interact closely with black patients and cannot diagnose them or advise them effectively. Maybe they are less willing to exert themselves and do everything within their power to help black patients. These are theories of *taste-based discrimination*. Alternatively, doctors might treat black patients differently because of biological differences between races or because the doctors believe black patients are less likely to follow up with medications or other doctor's orders and are not good candidates for treatment in the first place. This is an example of *statistical discrimination*. The important distinction from a welfare point of view is between *efficient* discrimination and *inefficient* discrimination.

Definition 5.6

Discrimination: Racial discrimination in medicine, defined as differential treatment rates for different racial groups, can be the result of several forces including racial animus (taste-based discrimination) and stereotyping based on biological differences (statistical discrimination). Discrimination can be medically efficient or inefficient.

Taste-based discrimination: preferential medical treatment for certain groups of patients due to the tastes of the physician, whether conscious or subconscious.

Statistical discrimination: differential medical treatment for certain groups of patients based on stereotypes of their biological or behavioral tendencies.

Efficient discrimination: differential treatment by race that improves patient outcomes on average.

Inefficient discrimination: differential treatment by race that harms patients on average.

Evaluating treatment disparities

One of many realms in which racial differentials have been noted is treatment decisions for acute myocardial infarction (AMI). Shulman et al. (1999) conducted a study of physician practice patterns to measure this differential. In this study, they invented fictional patient

histories involving angina pectoris, chest pain that may be an indicator of an impending AMI. The researchers then wrote an interview script for professional actors, some black and some white, to recite while being filmed sitting in a consultation room wearing a hospital gown. They created pairs of videos with identical transcripts, hand motions, and background props. In each pair of video clips, the only difference was the race of the "patient."

These patient history videos were shown to a sample of doctors, who were asked to make treatment recommendations for the hypothetical patients. The physicians, who were overwhelmingly white males, were significantly less likely to recommend black patients with chest pain for cardiac catheterization. This disparity arose despite the fact that the researchers designed the patient histories and interview transcripts to be identical across patients of different races.

This disparity in treatment is certainly evidence of discrimination – but what kind? The question remains whether the discrimination is taste-based or statistical and whether it leads to optimal or suboptimal outcomes for blacks and whites. Generally speaking, taste-based discrimination is inefficient because some factor other than a patient's welfare is affecting treatment decisions. Statistical discrimination, though, may be efficient if there is evidence that it makes medical sense to treat racial groups differently in some situations.

For example, medical research shows that optimal treatment for hypertension (high blood pressure) in black patients is different from the optimal treatment for whites. Gupta et al. (2010) argue that diuretics and ACE inhibitors should be used as first-line treatment for black hypertensive patients, while beta blockers and calcium channel blockers are the preferred treatment for white patients. In the case of hypertension, statistical discrimination actually saves lives.

Testing for inefficient discrimination

Chandra and Staiger (2010) attempt to determine whether the discriminatory prescriptions noted by Shulman et al. (1999) and others are efficient or inefficient using a subtle test. They analyze the medical records of black and white Medicare patients admitted to US emergency rooms suffering from AMI. They compare the rates at which black and white patients undergo procedures involving reperfusion (catheterization or open heart surgery). They find that black patients are slightly less likely to receive reperfusion than white patients, even after controlling other relevant risk factors. The differential reperfusion rates imply discrimination just like the Shulman et al. (1999) audit study.

Suppose this discrimination is the result of tastes, and also suppose that some patients can be expected to benefit more from reperfusion than others. Perhaps white doctors are willing to put in every effort to save white patients, and so recommend reperfusion even for white patients for whom it may only be marginally useful. Meanwhile, they only recommend reperfusion for black patients who will definitely benefit from it, and do not bother to treat marginal black patients who may not significantly benefit from reperfusion.

Under this hypothesis, the average benefit for black patients receiving reperfusion would be *higher* than white patients receiving reperfusion. Intuitively, this makes sense because, under this assumption, borderline white patients with bad prognoses are being included in the reperfusion group, while borderline black patients are not.

Surprisingly, the researchers find that reperfusion benefits black patients *less* than it benefits white patients; thirty-day survival rates among black patients are improved with

reperfusion, but the improvement is larger for white patients. This evidence is inconsistent with the hypothesis of discrimination against blacks and even suggests the reverse: physicians are either overtreating black patients or undertreating white patients. This finding makes clear a weakness of the audit study approach, because it shows that differential rates of treatment do not necessarily imply harmful discrimination.

5.6 Conclusion

Periodically we hear reports of shortages of certain types of doctors: primary care physicians, family medicine doctors, and – increasingly, with our aging population – geriatric care specialists (see Chapter 19). In free markets, the concept of a supply shortage has little economic meaning in the long run. If people are clamoring for more supply (that is, if demand is sufficient), new suppliers will enter the marketplace.

As we have seen, the labor market for physicians does not quite work this way. Physician supply does not respond quickly to demand changes because there are so many impediments to new suppliers entering in any given specialty. The training process for new doctors is long and arduous, and residency is so specialized that doctors cannot switch readily from specialty to specialty when the market demands it. Finally, significant barriers to entry play a role in diminishing the number of doctors available.

These training requirements and other barriers limit the supply of doctors, but they also benefit patients like the 59-year-old stroke victim who opened this chapter. She can be confident that her doctors have been trained to provide good care.

But doctors, as central as they are to the health care system, are not the only players who contribute to her health. Physicians coexist with hospitals, insurance companies, and pharmaceutical developers that also play their own role in health care provision. In the next chapter we examine the economics of the hospital industry and its recent evolution.

5.7 Exercises

Comprehension questions

Indicate whether the statement is true or false, and justify your answer. Be sure to cite evidence from the chapter and state any additional assumptions you may need.

1. Physicians in the US are licensed to practice medicine immediately after they complete medical school.
2. The internal rate of return is defined as the interest rate that makes the net present value of an investment stream exactly equal to zero.
3. Consider two investment streams w and z which pay out some amount, $w(t)$ and $z(t)$, in each period t. (The amount may be negative in some periods.) If the interest rate is exactly equal to the internal rate of return of $w(t)$, the net present value of choosing w over z is zero.
4. The number of US medical schools decreased drastically between 1900 and 1950.
5. The full economic cost of medical school includes mainly tuition, room, and board for the school.

6 Compared with doctors who are paid on a fee-for-service basis by health insurers, doctors who are paid on a capitated (per-patient) basis have incentives to provide too much care.
7 In part, physicians' salaries are higher than secretaries' salaries because it takes more years to train to become a physician than it does to become a secretary.
8 The fact that practicing surgeons who have finished residency earn more than practicing pediatricians implies that the rate of return of choosing surgery exceeds the rate of return of choosing pediatrics for a medical school graduate.
9 Once length of residency and hours of work are taken into account, the internal rate of return of choosing a specialized branch of medicine over a more generalized branch is roughly equal to the real rate of interest in the economy.
10 If physicians are earning monopoly rents, then there must be more barriers to entry in the labor market for physicians than is socially optimal.

Analytical problems

11 **Fun with IRR**. Suppose you have just graduated from college and are deciding on a career. Your four career options, along with your salary in each of the four earning periods, are displayed in Table 5.3. Assume that any career will only last four periods before retirement.

Table 5.3. *Career options and salary information for Exercises 11 and 12.*

	Salary			
Occupation	Period 0	Period 1	Period 2	Period 3
Ophthalmologist	−5	1	10	12
Accountant	2	3	4	5
Starving artist	1	1	1	1
Sports superstar	15	0	0	0

a Assume your discount factor $\delta = 0.95$. Interpret this assumption.
b Find the value of the interest rate r that corresponds to your discount factor.
c Assuming $\delta = 0.95$, calculate the net present value (NPV) of becoming an ophthalmologist and of becoming an accountant. Which career do you prefer?
d Will the internal rate of return (IRR) for becoming an ophthalmologist as opposed to an accountant be greater or less than your answer to Exercise 11(b)?
e Now assume $\delta = 0.6$. Calculate both the corresponding interest rate and the net present value (NPV) of becoming an ophthalmologist and of becoming an accountant. Now which job do you prefer?
f Find the IRR for becoming an ophthalmologist as opposed to an accountant. That is, find a value of r^* that equates these two NPVs. [*Hint*: you will probably want to use a graphing calculator or an online equation solver to find r^*.]

12 **More fun with IRR**. Refer to Table 5.3 about the payouts available at different jobs.
a Find the IRR for becoming an ophthalmologist as opposed to a professional sports star. Compare your result with the IRR from the previous exercise and interpret this difference in terms of the concept of patience.

b Find the IRR for becoming an accountant as opposed to an ophthalmologist. How can you interpret a negative IRR?

c The IRR for becoming an accountant as opposed to a starving artist is infinite. Explain why this makes sense.

d Does the NPV of salaries in the various professions tell you everything you need to know about picking a career? What does this calculation leave out?

13 In each of the following situations, indicate whether the physician's discriminatory action is taste-based or statistical.

a An American physician detests interacting with French people, so she always gives French patients quicker examinations than she gives Americans.

b A physician believes that Hispanic patients are less likely to follow through with an expensive therapeutic regimen that leads to major side effects, so he never prescribes it for them.

c A surgeon has heard that it is very difficult for black patients to find bio-compatible matches on the kidney donor waitlist. As a result, he is more aggressive in trying to save a black patient's kidneys than a white patient's kidneys.

d A hospitalist believes that the nurses at her hospital routinely discriminate against Asian patients due to subconscious racism. She always spends a little more time ensuring that her Asian patients are receiving the right medication.

Essay questions

14 A recent study by Prof. Jessica Reyes at Amherst College finds that female gynecologists (doctors who specialize in women's reproductive health) charge a standard fee for a basic patient visit that is $4.81 (on average) higher than male gynecologists charge. Furthermore, she finds that waiting lists to see female gynecologists are 1.14 weeks longer than waiting lists to see male gynecologists.

a Give two possible explanations for Prof. Reyes' findings. One explanation should focus on the demand side of the market for gynecologists; the other should focus on the supply side.

b Are the longer waiting times and higher fees for female gynecologists evidence that the market for gynecological services is not competitive? Give two answers to this question – one assuming that your demand-side explanation from Exercise 14(a) is correct, and one assuming that your supply-side explanation is correct.

c Has Prof. Reyes uncovered evidence that there is discrimination in the residency training market for gynecologists? If so, is this discrimination against male gynecologists or female gynecologists? Again give two answers to this question – one assuming that your demand-side explanation from Exercise 14(a) is correct, and one assuming that your supply-side explanation is correct.

15 Below is an excerpt from the abstract of a 2002 journal article entitled "Does regulation affect economic outcomes? The case of dentistry" by Morris Kleiner and Robert Kudrle:

> Theory suggests that more restrictive licensing may raise prices and at the same time raise demand by reducing uncertainty about the quality of the services. This article uses unique data on the dental health of incoming Air Force personnel to analyze empirically the effects of varying licensing stringency

among the states. It finds that tougher licensing does not improve outcomes.

a Explain the basic tradeoff a state makes when it tightens regulations on licenses for physicians, dentists, or any other professionals. What are the costs and benefits?

b Kleiner and Kudrle find that cadets hailing from states with tougher licensing regimes for dentists do not have healthier teeth on average. Argue that the tougher licensing regimes might be efficient, even if those states also have higher prices for dental services.

c Assume also that states choose strict or lenient licensure regimes at random (this is probably not the case), and that states with strict licensure have higher prices but no better dental health. Argue that the tough licensing regimes are inefficient.

d Why could you not conduct a similar study about licensing regimes for physicians using the same data on US Air Force cadets?

Students can find answers to the comprehension questions and lecturers can access an Instructor Manual with guideline answers to the analytical problems and essay questions at **www.palgrave.com/economics/bht**.

6 THE HOSPITAL INDUSTRY

Today the mention of a hospital invokes the image of a pristine fortress with clean, white rooms and futuristic technology. It is bustling with nurses, physicians, surgeons, other specialists, and administrators taking care of patients and moving from task to task with clockwork efficiency. A hospital is a place where the sick seek state-of-the-art treatments for whatever ails them.

But today's hospitals would have astonished a visitor from the mid-1800s. In that era, hospitals were generally viewed as squalid places where only the most indigent, desperate members of society sought care. The density of patients packed into unkempt wards coupled with a primitive understanding of germ theory led to higher mortality rates after surgery in hospitals than in homes. Unsurprisingly, members of the upper and middle classes did their best to avoid the hospital, and since most doctors made home visits or held private practices, avoiding the hospital was never all that difficult for those who could afford it (Starr 1982).

Several innovations arrived in the late 1800s to fundamentally transform hospitals. The introduction of anesthesia turned surgery from brutal and risky to life-saving and humane. Aseptic techniques motivated by a better understanding of germs ensured that surgeries proceeded only in sterile environments. The invention of X-ray technology in the 1890s further enabled doctors to diagnose and treat disease and injury.

The dramatic increase in the power of surgery also demanded dramatically more resources – surgeons needed sterile rooms in which to operate, the support staff to conduct these operations, and beds to monitor patient recovery. By the early 1900s, hospitals were becoming the places for medical care, and the earlier stigma associated with hospitals was starting to fade.

With higher costs, hospitals also began developing controversial ways of funding themselves. In the early 1900s, most hospitals were affiliated with religious institutions, and most refrained from charging fees to poorer patients who stayed in densely packed wards rather than private rooms. Some hospitals operated by soliciting donations from altruistic patrons, others by charging richer patients for staying in private rooms. Modern hospital financing looks very different. Insurance systems are responsible for almost all payments. Increasingly, hospitals are adopting for-profit models and eschewing the traditional nonprofit model of religious charity.

6.1 The rise and decline of the modern hospital

The Hill-Burton Act and the rise of the hospital

Scientific progress, such as the discovery of penicillin and other advances, transformed the hospital from simply a place to die into a vital input into improving the health of a population. By the middle of the twentieth century, there was growing concern that

there was an insufficient number of hospitals in the US to adequately care for everyone, especially in rural areas. Furthermore, the rising costs of care put hospitals increasingly out of reach of the poor, even in cases of medical emergencies.

In 1946, the US Congress passed the Hill-Burton Act which was designed to address both of these problems. The Act provided substantial monies for the building of hospitals around the country with a preference toward underserved rural areas. Furthermore, the Act mandated that any hospital accepting Hill-Burton money would be required to provide free or low-cost care for the poor and indigent. Federal funding and the rise in demand together inspired a spree of hospital building around the country which continued more or less unabated until 1974 (Figure 6.1). The number of hospitals increased by 16% and the total number of hospital beds by 12% during this period.

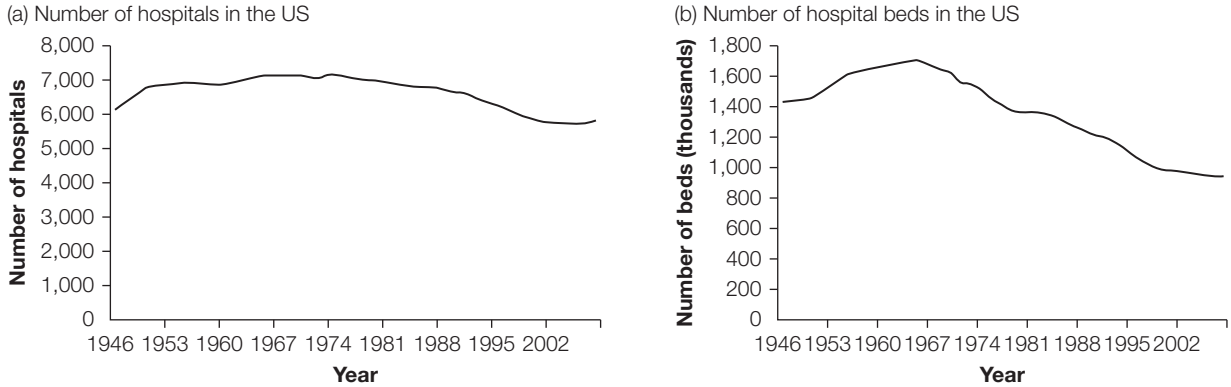


Figure 6.1. *Recent trends in US hospitals. (a) Number of hospitals in the US. (b) Number of hospital beds in the US.*

Source: Data from American Hospital Association (2010).

The transition to outpatient care

In 1974, the number of hospitals peaked and has been declining ever since. In fact, the probability of being hospitalized in any given year has dropped sharply since then as well. Even as the number of hospitals and hospital beds was growing in the mid-1900s, the average length of a hospital stay has steadily declined since 1946 (Figure 6.2). In American hospitals, average lengths of stay are the shortest in the developed world. Hospitals are still important providers of inpatient care for the severely ill and they continue to generate enormous expenditures. However, hospitals no longer play the predominant role in providing medical care. Instead, a large portion of medical care now takes place outside the hospital, in outpatient clinics.

There are at least two important reasons for the decline of inpatient care, one technological and one economic.

Many of the scientific advances in medical care over the past decades have made it possible to perform services that previously required overnight hospital care in outpatient settings; hospitals are therefore less necessary than they once were. One good example of technological change is the development of laparoscopic surgery. In laparoscopic surgery, surgeons rely on a small camera, rather than a gaping incision, to see into the patient's body. It came into widespread use during the 1980s and 1990s and has greatly shortened the recovery time for many surgeries.

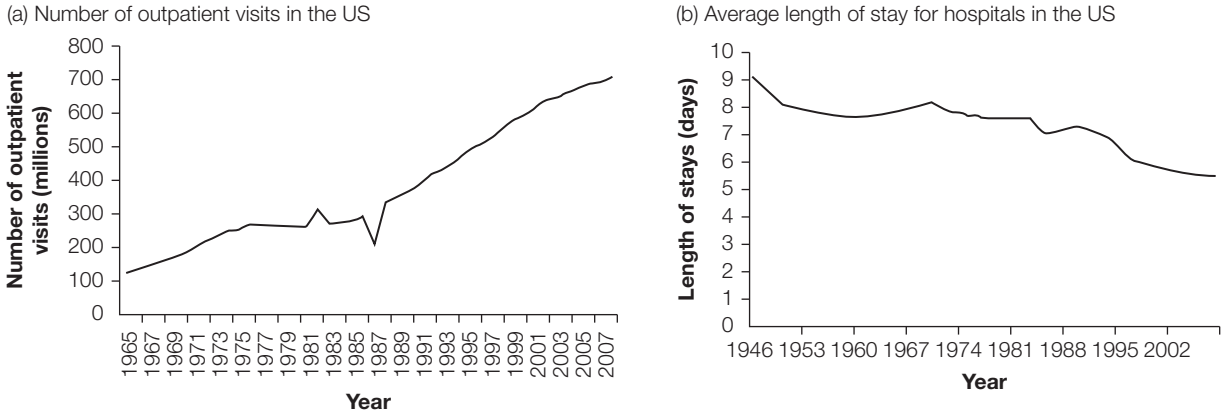


Figure 6.2. *Recent trends in US hospitals. (a) Number of outpatient visits in the US. (b) Average length of stay for hospitals in the US.*

Source: Data from American Hospital Association (2010).

Cholecystectomy, an operation to remove the gallbladder, historically required weeks of hospitalization for recovery. The traditional method, an open cholecystectomy, required a deep incision from the middle of the abdomen to the right hip. A long hospital stay was necessary for doctors to monitor recovery and to treat any infections. But since the advent of new laparoscopic techniques, a gallbladder can be removed with just three small, strategically-placed incisions in most cases. This less invasive surgery typically no longer requires an overnight stay in the hospital (Ahmad et al. 2008).

A second reason for the transition to outpatient care is that insurers have changed the way they compensate hospitals. These changes have made hospital stays, especially long ones, less lucrative for hospitals, particularly in the US. In 1984, the US Medicare program, which provides universal health insurance coverage for the elderly, introduced a prospective payment system for hospital reimbursement known as the diagnosis-related groups (DRG) payment system (Cutler 1995). Prior to this change, hospitals were paid on a fee-for-service (FFS) basis for each day that a patient stayed in the hospital and for each service performed. Since fees were set above marginal cost, hospitals had financial incentives to provide even marginally necessary services and extend hospital stays. Before the DRG system, Medicare faced all the financial risks associated with ongoing hospital stays.

Under the DRG system, hospitals receive a fixed fee that depends on the patient's diagnosis at the time of hospital admission and on nothing else. In particular, hospitals do not receive extra payment if the patient stays in the hospital for a long time or receives costly services. Unlike with the FFS system, the hospital bears the majority of the financial risk associated with long hospital stays. Perhaps unsurprisingly, the introduction of the DRG system preceded a sharp drop in the average hospital length of stay in the US (see Figure 6.2). In subsequent years, many European health systems with private hospitals have moved towards a DRG system.

In contrast with the US, Japanese hospitals are still paid on an FFS basis (though recent reforms have introduced a partial and voluntary system of prospective payment). Though Japan's average length of stay has declined over the past decades, presumably because of advances in medical technology, it is still substantially higher than the average length of stay in the US. Under Japan's fee-for-service payment system, hospitals have little incentive to avoid lengthy hospital stays (see Chapter 17).

Despite all the technological advancements of the last century, hospitals remain dangerous places. They are still places where sick people are housed in close quarters and can

spread diseases to one another. Hospitals are also breeding grounds for drug-resistant bacteria that threaten the health of the elderly and the immuno-compromised. Given these considerations, the recent trend away from inpatient care to outpatient care is welcome.

6.2 The relationship between hospitals and physicians

Today, the modern hospital combines the great expertise of medical professionals with expensive technologies. Running a large hospital is a coordination challenge that requires extensive manpower for tasks such as managing the operating room, conducting blood tests and X-rays, and dispensing drugs.

People tend to think of "good hospitals" as ones with excellent physicians who are well trained and have decades of experience in their various medical specialties. But while the physician may be the face of the hospital in the patient's mind, it is a vast support staff of nurses, orderlies, clerical workers, and executives who keep the hospital running. We will see that hospitals are organizations that can learn from experience just as physicians do. A good hospital is not necessarily the one with the best doctors, but the one whose doctors and support staff work together best.

The internal organization of hospitals

Harris (1977) argues that hospitals should be thought of as two separate economic entities: the physician staff and the administrative staff. The physician staff treats patients and demands medical goods and services like syringes, MRI tests, operating rooms, and nursing care. The administrative staff consists of nurses, executives, and other hospital employees who work to supply these inputs demanded by physicians. The physician staff and administrative staff face different incentives but are together responsible for the decisions made by hospitals, such as the adoption of new technologies.

Unlike the administrative staff, physicians are often not direct employees of the hospital. In the US, physicians may run their own private practices but refer their sickest patients to hospitals. Then as part of their relationship with the hospital, the physicians can use its resources to continue treating their patients. The patient or his insurance continues to pay the physician and pays the hospital separately for its facilities. The administrative staff does not interfere with the doctor's treatment of patients. Given the power that doctors often wield in hospital settings, hospitals have long been known as the "physicians' workbench."

The fact that doctors are not employees of the hospitals tends to undermine their incentive to control hospital costs. Instead, doctors plead with hospital administrators to adopt every technological innovation to treat their patients, even if they are expensive or useless. The hospital administrators control the budget but typically lack the medical knowledge to judge the relative value of different innovations for patient health. This bargaining process between doctors and administrators can result in wasteful spending to install cost-ineffective technology. This allocation of decision power has been cited as a possible explanation for why hospital costs have grown so much in past decades.

In practice, there are three different models of the relationship between physicians and hospitals. The first is the "physician's workbench" relationship, described by the Harris model, in which hospitals provide a place for physicians to do their work, but do not directly employ them. The Harris model describes most American hospitals reasonably well.

The standard example of learning comes from a pin factory described in Adam Smith's Wealth of Nations *(1776). Specialization of labor within the factory enabled workers to become better at a very specific task – like straightening a wire or sharpening the point of a pin – through repeated practice.* *Credit: © Georgios Kollidas – Fotolia.com.*

The second type of relationship between physicians and hospitals involves physicians as direct employees of a hospital or hospital system. In the UK, for instance, most physicians are employees of the National Health System, which runs most of the nation's hospitals. This type of arrangement is becoming more popular in the US as well, even at private hospitals.

The final type of relationship involves physician ownership of hospitals, which is traditionally justified as a way to avoid the commercialization of medicine (Rodwin and Okamoto 2000). In Japan, for example, it is a common for a physician or a small group of physicians to own small hospitals and clinics.

These alternate types of relationships avoid the conflict of interest between physicians and administrators. But because doctors must also be mindful of costs, there is the potential that doctors will fail to serve as ideal agents for their patients. For instance, sometimes doctors may choose to save the hospital money rather than exploring every possible avenue of treatment.

The volume–outcome relationship and learning by doing

Within hospitals, economists have long noticed a positive correlation between the number of cases and the outcome of those cases. For example, surgeons and hospitals who perform the same procedure over and over again seem to have better post-operative results and fewer complications. This relationship between volume and outcome is called the **volume–outcome relationship**.

Definition 6.1

Volume–outcome relationship: the observed positive correlation between the number of procedures performed (volume) and patient outcomes.

One hypothesis for the volume–outcome relationship is *learning by doing*, the process whereby people get better at doing tasks they perform repeatedly. To what extent can the volume–outcome relationship be attributed to learning by doing? It certainly seems plausible for surgical teams who complete a complex procedure nearly every day to outperform teams at hospitals who complete the same procedure only a few times a year.

On the other hand, the volume–outcome relationship may be the result of selective referral patterns: doctors may prefer to direct their patients to the most highly skilled surgeons. Even if there is no learning by doing, this selective referral hypothesis would explain the volume–outcome relationship. The best surgeons, who tend to have the best patient outcomes, receive the most referrals, and as a result, a positive relationship emerges between volume and outcomes. Under this alternate hypothesis, the direction of causation is reversed from learning by doing: variation in outcome causes variation in volume and not the other way around. Both the learning-by-doing hypothesis and the selective-referral hypothesis seem plausible, and evidence has been found in support of both (Luft et al. 1987).

Recent studies have highlighted the fact that the strength of the volume–outcome relationship in hospitals varies dramatically across different procedure types. Birkmeyer

Table 6.1. *Surgical mortality rates for various Medicare procedures, by hospital volume.*

	Hospital volume		
Procedure	Lowest 20%	Middle 20%	Highest 20%
Coronary-artery bypass grafting[a]	6.1	5.3*	4.8*
Aortic-valve replacement[b]	9.9	9.1*	7.6*
Carotid endarterectomy[c]	2.0	1.8*	1.7*
Pancreatic resection[d]	17.6	11.6*	3.8*
Nepherectomy[e]	3.6	2.7*	2.6*

[a] Coronary-artery bypass grafting (CABG) involves replacing one or more clogged coronary arteries, which provide blood to the heart muscle, with a vein typically taken from the thigh.
[b] The aortic valve mediates blood flow between the heart's left ventricle and the aorta, which is the biggest artery in the body. It can be replaced with a mechanical valve or an aortic valve from a pig or a cow.
[c] Carotid endarterectomy involves the removal of atherosclerotic plaque from the carotid artery in the human neck. The surgery requires partially blocking blood flow to the brain, so it must be completed quickly.
[d] The removal of part of the pancreas is a complex surgical procedure that typically takes several hours to complete. Cancer is a common reason why pancreatic resections are performed.
[e] Nepherectomy is the removal of the kidney: *-ectomy* means removal, and *nepher-* means kidney.
* Indicates statistically significant difference from hospitals with the lowest 20% of volumes.

Source: Data from Birkmeyer et al. (2002).

et al. (2002) use Medicare data from the 1990s and find evidence of volume–outcome relationships for all 14 procedures they study. A selection of their results is shown in Table 6.1. Even in the sample, substantial variation exists: the correlation is stronger for pancreatic resection than it is for carotid endarterectomy. Meta-analyses of volume–outcome studies similarly find wide variation by procedure type (Halm et al. 2002).

Hospital experience versus physician experience

It is clear that *hospital* volume is correlated with outcomes. It is also true that *physician* volume is correlated with outcomes. Which one is more closely related? A recent study compares the volume–outcome relationship for individual physicians with the volume–outcome relationship for hospitals. McGrath et al. (2000) study the outcomes of Medicare patients undergoing percutaneous coronary intervention (PCI) to unclog their coronary arteries, tracking both 30-day mortality and the rate of follow-up coronary artery bypass graft (CABG) surgery. If a patient needs CABG surgery soon after a PCI, it is an indication the PCI failed to keep the coronary artery unclogged, so the need for CABG can be seen as evidence of an ineffective PCI.

In both cases, patients experience significantly worse outcomes when served by low-volume providers compared with high-volume providers. But as is apparent in the data, the volume–outcome relationship for hospitals is greater in magnitude. For instance, an increase in hospital PCI volume from 75 to 200 annually results in a decline in 30-day mortality from about 4% to 3%, while a similar increase in physician PCI volume is not associated with any decrease. In the case of PCI, the experience of the hospital seems to be more important than the experience of the attending surgeon.

The finding that hospital experience matters so much might be surprising, but it makes sense as surgeries are complicated exercises that require the coordination of several hospital employees, not just the leading physician. A typical procedure involves not only the surgeons, anesthesiologists, and nurses present in the operating room, but also pharmacists and operating assistants who gather the tools and information for the doctors.

And after the surgery, a separate post-operative team might be managing the patient's recovery.

Errors may occur before, during, and after a surgery. For instance, before a cataract surgery on a patient's left eye, the surgeon might mistakenly mark the patient's right eye for operation. During the actual surgery, the anesthesiologist might overdose the patient. And after the surgery, the team caring for the patient might not have been told about the patient's diabetes, resulting in a medication error. De Vries et al. (2010) estimate that more than half of all surgical errors actually occur outside the operating room.

A key to preventing these errors is often to implement systems at the hospital level that require multiple people to double-check each step of the patient's care. De Vries et al. find that hospitals perform significantly better when employing a checklist system designed to promote communication between the different hospital employees involved in any surgery. Given the coordination across the hospital that is necessary to care for a patient, it is not surprising that organizations appear to gain from learning by doing as much as or more than individual doctors do.

The rise of the hospitalists

Despite the well-documented benefits of learning by doing, few doctors in the US specialized exclusively in hospital care before the mid-1990s. Such specialists, known as **hospitalists**, had existed in Canada and the UK before this time, but almost all hospitalized patients in the US were cared for by their primary care doctor, rather than an in-house physician specializing in hospital care (Wachter and Goldman 1996).

Since the mid-1990s, the hospitalist model has experienced tremendous growth in the US. These hospitalists perform the same types of procedures as primary care doctors, but only treat patients who have been hospitalized. Several reasons have been proposed for this change, including cost pressures on physicians and hospitals, increased time burdens on primary care physicians, and the declining role of inpatient care within the medical system (Wachter and Goldman 2002). The evidence on learning by doing in medicine may also be motivating this shift. There is much potential for learning by doing by hospitalists, who typically spend six months or more of the year in hospital settings. Their experience may make them more efficient at treating hospitalized patients than non-hospitalists who only treat inpatients sporadically.

A meta-analysis of 33 studies published between 1998 and 2008 shows that hospitalists usually outperform their non-hospitalist counterparts, such as community doctors and general internists, in terms of reducing patient length of stay and lowering total medical expenditures at hospitals (Peterson 2009). A few studies also show better mortality outcomes for patients treated by hospitalists. Most of these studies feature randomized assignment of patients to hospitalists and non-hospitalists, so they provide strong causal evidence of the benefits of treatment by specialists.

Another hypothesized benefit to having hospitalists care for inpatients is that hospitalists tend to be more rapid and efficient adopters of new technologies. One possible mechanism is that hospitalists have greater incentive to stay abreast of advancements in the care of inpatients, because that is their focus. This willingness to try new technologies can benefit other patients in the same hospital. Because hospitalists are around the hospital continuously, they are more likely to develop social networks along which knowledge can diffuse (Meltzer 2009).

Health economist and physician David Meltzer relates the story of the adoption of low-molecular-weight heparin (LMWH), a drug used to treat stroke patients, at the University of Chicago Medical Center. A single hospitalist learned about the drug and was the first to use it on a patient at the center. Other doctors, including non-hospitalists and trainees who were on the team caring for that patient, observed this and subsequently rotated onto new teams in the months that followed. They, in turn, introduced LMWH to their new colleagues, further spreading knowledge of the drug throughout the hospital.

6.3 The relationship between hospitals and other hospitals

Like suppliers in other industries, hospitals compete with one another for customers – in this case, patients. But several characteristics about the market for health care make hospital competition dissimilar from competition in other industries.

First, in some countries, access to health care is seen as a basic right, and hospitals are forbidden from denying anyone care. Second, building and staffing a new hospital requires high initial investment costs and often some form of government approval. These high barriers to entry can impede competitors from entering the market. Lastly, the ubiquity of insurance in this market distorts the typical economics of supply and demand. As a result of these quirks in the hospital market, the standard analysis of competition must be amended.

Differentiated product oligopoly

Many economic theories start with the assumption that markets are perfectly competitive. While no market is truly *perfectly* competitive, the market for hospitals strays further away from this ideal than most. For one, entry is not easy. New hospitals have to invest substantial funds in erecting buildings, installing technologies, and hiring physicians and administrators. Governments also regulate the entry of health care providers. In most US states, firms are required to obtain a Certificate of Need approval from the state government before building a new hospital or expanding an existing one. These barriers restrict entry into the hospital market, so there are often only a few hospitals competing for patients in a given geographic area. Economists typically model these markets as *oligopolies* (Gaynor and Town 2013).

Additionally, hospitals are not perfect substitutes for one another. Each hospital may appear superficially similar – they all have operating rooms and radiology labs – but in fact they differ on both the set of services they offer and the quality of each service.

The type of care a heart attack victim receives in the emergency room may depend on whether the hospital is equipped with a catheterization lab (McClellan et al. 1994). Even two hospitals with catheterization labs may provide different quality of care depending on which hospital's staff and doctors are more experienced.

And even if the standard of medical care is exactly equal across hospitals, patients may have some loyalty to their own physician or a surgeon they worked with in the past. Patients also live different distances away from different hospitals. Especially in an emergency, travel time can differentiate the services of each hospital.

Because of the barriers to entry and the differences between suppliers, economists model the hospital market as a **differentiated product oligopoly**.

Definition 6.2

Differentiated product oligopoly: a model of competition in which there are few firms as a result of barriers that restrict entry, and in which the products supplied by the firms are not perfect substitutes for each other.

Suppose that the marginal cost of performing a hip replacement surgery is $10,000. This includes the cost of all the resources necessary for the operation: alcohol swabs, a surgeon's time, medical imaging equipment, and everything else. Further suppose an avaricious hospital tries to make more money by charging $11,000 for each hip replacement surgery it conducts. In a perfectly competitive market, this move would be economically suicidal: a new hospital would enter the market, charge $10,500 for the same procedure, and attract every last customer from the first hospital. Other firms would enter until the price stabilizes at $10,000.

But this story relies on two critical assumptions: easy entry of new firms and the identicalness of goods. As we have discussed, neither assumption holds in a differentiated product oligopoly. As a result, a hospital in an oligopoly might get away with charging $11,000 without losing all its customers. Some customers may stay because they feel a loyalty to their longtime hospital, believe that the hip replacement offered by this hospital is superior, or are unwilling to travel to more distant hospitals for a discount. Furthermore, it may be infeasible for an enterprising firm to build a new hospital to undercut the expensive one.

Because oligopolistic firms are not simply price-takers, but can instead price above marginal cost and earn positive economic profits, they are said to have *market power*. Firms that have market power face downward-sloping demand – unlike firms in perfectly competitive markets, a firm with market power can raise prices without losing all its customers. Market power is undesirable for society because it leads to high prices for consumers, less productive output, and lower social welfare.

Oligopolistic firms wield some market power, but their power is limited because, unlike monopolists, they do face some competition. Suppose the same hospital decided to charge $1,000,000 for each hip replacement surgery in an attempt to leverage its market power. Now many or all of its patients would probably leave, even if they live far from other hospitals or feel an intense loyalty to their doctors.

Because there are few firms in oligopoly markets, it might become possible for all firms to cooperate, agree to maintain high prices together, and split up the available profits. This practice, called *collusion*, is forbidden by governments because it exacerbates the social welfare loss from oligopoly. But it still might happen illegally if firms can communicate in secret or tacitly signal their intentions to each other through prices. Strategic behavior by oligopolists can also be non-cooperative. For example, incumbent firms may temporarily lower prices below marginal cost to bankrupt competitors and extend their market power. Strategic pricing behavior cannot happen in perfectly competitive markets, since prices are constrained to equal marginal cost.

Oligopoly markets fall somewhere between perfectly competitive markets and monopolies. Oligopolies cause the same problems that monopolies do (high prices and low output), but due to the presence of limited competition, these problems are not as severe with oligopolies as they are with monopolies.

To determine whether hospital markets are oligopolies as theory suggests, economists use a measure of market concentration called the **Herfindahl–Hirschman Index** (HHI),

which ranges between 0 and 1. A low HHI indicates that a market is fairly competitive, with customers spread out evenly across many firms. A high HHI means that a small group of firms dominates the market, and may indicate an oligopolistic market.

Definition 6.3

Herfindahl–Hirschman Index: a measure of market concentration. It is the sum of the squared market shares of all of the firms in a market:

$$HHI = \sum_i s_i^2$$

where s_i is the market share of firm i. The highest possible HHI is 1, indicating that one firm controls the entire market. Theoretically, the lowest possible HHI is 0, indicating an infinite number of firms with infinitely small market shares.

The US Department of Justice and the Federal Trade Commission, which share responsibility for regulating competition in markets, label a market as "concentrated" if its HHI is greater than 0.18, and "highly concentrated" if its HHI is greater than 0.25. By this measure, hospital markets in the US are not very competitive: the HHI for the hospital market in the average American metroposition area was 0.33 in 2006 (Gaynor and Town 2013).

This evidence suggests that hospitals do indeed have some market power and there may be too little competition in hospital markets. If so, policies aimed at increasing competition will lead to lower prices and better outcomes for patients. But hospital markets have another unusual property: incomplete price competition. Unlike nearly all other markets, markets with this property do not necessarily benefit from more competition.

Price competition

In most markets, suppliers attempt to attract customers by offering lower prices than their competitors. In perfectly competitive markets, and even to some extent in oligopolistic markets, consumers punish firms that charge high prices by taking their business elsewhere. If a drug store decides to start charging $100 for a generic painkiller, customers would quickly abandon that store and buy their painkillers elsewhere. Price competition keeps markets efficient by forcing firms to offer their clients good value or face the risk of bankruptcy. But when price competition is limited, this check on inefficiently high prices may disappear.

Insurance, which is ubiquitous in health care markets generally and hospital markets specifically, can interfere with price competition and may eliminate it completely. For example, patients in the US covered by Medicare pay only a fixed amount per hospital visit, and medical costs in the UK are completely covered by the National Health Service (NHS). Thus, insured patients have diminished incentive to search for the lowest-cost providers. If most of the burden of higher medical prices is shouldered by the insurer, then patients may be indifferent between expensive and inexpensive hospitals. This distortion of incentives for insured patients is known as *moral hazard*, a phenomenon that appears again and again in health care markets (see Chapter 11).

Moreover, search costs are often high and price information can be hard to come by. In emergencies such as a heart attack, the patient is rushed to the nearest emergency room,

without concern for price. And in most medical situations, the full price of treatment is usually uncertain, depending on a yet-unknown diagnosis and various complications that arise during treatment.

In many health care systems, medical prices are actually set by the government. For example, Medicare and Medicaid in the US determine at the beginning of each year how much it will pay for each group of treatments for a subset of the American population. Similarly the UK's NHS determines the payment for hospital care. In these cases, hospitals cannot lower prices to attract more patients. But more generally, even when payers are not unilaterally setting hospital prices, the presence of insurance always undermines price competition.

Price competition is not totally absent in hospital markets though. For instance, any patient who must cover some of the costs of her own care – usually in the form of a copayment – still has some incentive to avoid pricey hospitals. And insurers themselves also foster price competition by directing patients to less expensive hospitals. For instance, managed care organizations (MCOs) in the US operate by restricting enrollees to a list of approved health care providers. MCOs often exclude expensive hospitals from their lists to maintain lower costs.

Quality competition

What happens when hospitals are no longer forced to compete on price? They tend to raise prices and compete with each other on quality instead. Quality is a broad concept that encompasses many aspects of a hospital visit, including the comfort of hospital beds, the availability of advanced technologies, and even the effort or vigilance of the staff. If a hospital can get away with charging high prices, it can attract patients with lavish, high-quality care and then charge them or their insurers a fortune for their stay.

The higher quality also serves to convince physicians to refer their patients to those hospitals. Hospitals offering the highest quality attract the most patients, while hospitals with the least lose patients, money, and may even have to shut down. When there are many hospitals competing for the same set of customers, a never-ending race for the best medical technology can develop. This is known as the **medical arms race hypothesis** (Dranove and Satterthwaite 2000).

Definition 6.4

Medical arms race hypothesis: hospitals compete on quality by adopting the best medical technologies available to appeal to physicians and their patients. This competition results in a race for each hospital to have the best medical technologies available and may cause overconsumption of medical technology.

If the medical arms race hypothesis is true, then competition among hospitals can be bad for society. The more hospitals compete with each other, the more frenzied the arms race becomes. Each hospital purchases expensive machinery, and redundancy occurs as multiple hospitals invest in the same fixed costs. Moreover, if physician-induced demand is present, then access to medical resources makes the use of such resources more frequent, even if that use is extraneous. So the excess technologies may also increase consumption, and health care expenditures will rise without corresponding improvements in health outcomes.

Empirical investigations into the medical arms race hypothesis have returned mixed findings. Most studies use positive correlations between competition in an area and the costs per admission as proxy evidence for a medical arms race. A hospital competing with many other hospitals in the area would invest more in medical resources and also generate higher expenditures per patient. Indeed, Robinson and Luft (1985) find that US hospitals in more competitive markets have higher expenditures. These hospitals also tend to have higher ratios of employees to patients, more angioplasty and coronary bypass facilities, and more mammography, heart surgery, and catheterization services (Robinson 1988; Robinson and Luft 1987; Luft et al. 1987).

Yet, the extent of the medical arms race might have shifted over time as a result of changes in the competitive landscape. For example, the increased popularity of MCOs during the 1980s in the US might have tamped down the medical arms race. Because MCOs vertically integrate the payer and provider of health care, they are sensitive to both hospital costs and patient bills. This sensitivity may limit any medical arms race if hospitals are more reluctant to invest in expensive technologies. Zwanziger and Melnick (1988) studied California hospitals between 1983 and 1985 and found that the positive correlation between competition and costs found in 1983 disappeared by 1985. Connor et al. (1997) similarly found that the correlation for hospitals nationwide greatly subsided by 1994 (see Chapter 18 for more on MCOs).

If the US medical arms race has cooled, then more hospital competition may not increase social loss. Nonetheless, Devers et al. (2003), after interviewing hospital executives in 1996 and again in 2001, have found that hospitals may be reviving their non-price competition strategies. The authors attribute this change to the waning popularity of MCOs and the rise of outpatient clinics that also compete with hospitals for patients (Berenson et al. 2006). If so, a medical arms race may be brewing again.

Only rarely does the medical arms race devolve into open medical warfare.

Credit: Allen Cox.

However, Dranove and Satterthwaite (1992) argue that even those studies that find positive correlations between competition and per-patient hospital expenditures may not actually be evidence of a medical arms race. They propose an alternative explanation for the correlation. More hospitals within a region can arise as a response to greater demand there, and that higher demand alone can produce a positive correlation. Meanwhile hospitals in smaller markets face less demand and therefore less return to installing MRI machines, angioplasty facilities, and other technology investments. As a result, these hospitals are unlikely to adopt the new technologies as quickly.

Thus, a positive competition–cost correlation could be the result of differing regional demand and not a medical arms race. Dranove and Satterthwaite test their hypothesis by applying an empirical strategy that controls for traits of different regional markets, and they find little support for the medical arms race hypothesis, even before 1983.

Hospital competition and patient outcomes

Thus, the standard prediction that more competition will improve welfare in the face of oligopolistic pricing may not hold in this market. Under certain conditions, increased hospital competition may actually lower consumer welfare by worsening patient outcomes. The exact effect of competition can depend greatly on many variables in the market environment (Gaynor and Town 2013).

Kessler and McClellan (2000) study the impact of hospital competition on all non-rural Medicare patients in the US treated for acute myocardial infarctions (AMI). They compared patient outcomes and medical costs for treatment in regions with more competition and regions with less. At least since 1991, more competitive regions seem to yield better results, including both lower costs and lower patient mortality. This suggests that more competition improves welfare for AMI patients.

Gowrisankaran and Town (2003) also study AMI patients on Medicare but focus on a specific geographical area in California: Los Angeles County. Unlike Kessler and McClellan, they find worse mortality rates in the more competitive parts of the county. This difference may have arisen because the first study used nationwide data, whereas Gowrisankaran and Town examined only one small region. Together, the studies suggest that the relationship between hospital competition and patient outcomes is not clear.

Even in markets with private insurance, the effect of competition on patient outcomes is unclear. Escarce et al. (2006) track mortality for six conditions: AMI, hip fracture, stroke, gastrointestinal hemorrhage, congestive heart failure, and diabetes. They find that higher competition led to better mortality outcomes for five of the conditions in California and for all six conditions in New York. In contrast, mortality did not vary with the level of competition for any of the six conditions at hospitals in Wisconsin.

In the UK, most hospitals are government-run. Nonetheless, various government reforms have tried to inject some elements of competition into the market. For example, some reforms in the 1990s permitted competitive bidding by hospitals for government contracts. Others in the 2000s permitted patients to choose which hospitals they visited for surgery and granted more autonomy to hospital administrators.

Two studies focus on NHS reforms in the 2000s that permitted greater patient choice in hospitals and more autonomy in hospital administration, even while health care prices remained centrally decided by the government (Gaynor et al. 2010; Cooper et al. 2011). Both find that mortality rates from AMI were lower in markets in more hospital competition. However, Propper et al. (2008) studying the 1990s reforms find that competition reduced waiting times, but also raised patient mortality rates.

Learning by doing may also explain why greater competition among hospitals may hurt patient outcomes. When many hospitals compete in a local area for a fixed number of patients, the number of patients treated at each hospital is low, so opportunities for learning-by-doing are also few. Meanwhile, a monopolistic hospital may charge high prices, which lowers patient welfare, but may gain invaluable experience through access to a large number of patients, which benefits the local community and more than compensates for the high prices (Lakdawalla and Philipson 1998; Gaynor and Town 2013).

The ambiguous relationship between competition and patient outcomes has motivated all nations to regulate the hospital industry. Certain governments have even taken the step of nationalizing their hospital industries and eliminating private competition altogether. We discuss more specifically the merits and drawbacks of different policy choices in Chapter 15.

6.4 Nonprofits and hospital production

Another motivation for public involvement in the provision of health care is equity. The notion that everyone should have access to basic health care underlies laws in the US

that forbid any hospital from denying emergency care to anyone for any reason, including inability to pay. In the UK, nearly all hospitals are government-run and publicly financed. Medical care at these hospitals is completely free to all citizens. In the US, public hospitals do charge their insured customers, but they offer free services to the indigent and usually provide the less lucrative services that private hospitals sometimes avoid (Horwitz 2005).

However, there is only a small percentage of public hospitals in the US, and these are typically run by state and local governments. In 2009, about 20% of American hospitals were publicly run. This may be one reason for the major presence of nonprofit firms in America's hospital industry. In the US, 75% of private hospitals were nonprofit in 2009 (American Hospital Association 2010).

In this section, we discuss the costs and benefits of nonprofit organization, as well as four prominent theories that seek to explain why nonprofits exist at all. The first three theories posit that altruistic preferences, government failure, and distrust of for-profits explain the existence of nonprofit firms. The final theory argues that nonprofits and for-profits are not all that dissimilar and that nonprofit status is an accounting technique motivated by a desire to evade corporate income taxes.

The costs and benefits of nonprofit organization

Unlike for-profit hospitals, nonprofit ones are exempt from certain taxes and are typically better able to attract donations because the government considers those contributions tax-deductible. However, in exchange for these benefits, nonprofits are forbidden from selling stock to raise capital or formally distributing profits to owners. Table 6.2 summarizes these costs and benefits, and Table 6.3 lists the four theories for the existence of nonprofits that we discuss next.

Table 6.2. *The costs and benefits of nonprofit status under US law.*

Benefits of nonprofit status	*Costs* of nonprofit status
• Exempt from corporate income taxes	• Forbidden from selling stock to raise capital
• Donors receive tax deduction for donations	• Cannot formally distribute profits to owners
	• Restricted to certain charitable activities

The government-failure theory

The government-failure theory of nonprofits predicts that nonprofit firms arise when the political process fails to provide as much of a charitable good or service as some members of the community would like to see. Some people gain utility if they know that homeless people in their city are receiving free medical care or if the local performing arts are thriving, even if they do not receive any direct benefits. In this case, the government responds to this altruistic sentiment by providing some subsidies for medical care and support for theater companies. Those people most interested in charitable services, though, may not consider these government efforts sufficient (Weisbrod 1975).

Mother Teresa was a Catholic nun who established orphanages in India and became world-famous for her ministry to the poor. She was outspoken in her belief that governments and other institutions fail to care for the poor. *Credit: © zatletic – Fotolia.com.*

Some might donate their time and effort individually to make up the shortfall in charitable services provision, but many people might lack the requisite time or skills to contribute. Under this theory, a nonprofit firm is a vehicle to organize the labor and capital of the people who demand more charity. As long as donors are willing to support altruistic causes above and beyond what the government provides, nonprofit firms will exist to satisfy this demand.

The altruistic-motive theory

Alternatively, nonprofits may arise as a reflection of the goals of altruistic owners (Newhouse 1970; Rose-Ackerman 1996). Consider a hospital run by a board of trustees. The board wants the hospital to succeed financially, but may also be concerned with the quantity and quality of medical services provided, especially to individuals who cannot afford them at market rates. They are willing to accept reduced profits in exchange for more output or higher-quality output. If a firm cares about something other than profit, it is not behaving like the typical profit-maximizing firm that economists study.

A firm that is not interested solely in maximizing profit may organize as a nonprofit because of the costs and benefits outlined in Table 6.2. Nonprofit legal status can help the firm increase output by reducing tax liability and attracting donations. These extra funds can allow nonprofit hospitals to build a new wing, add extra beds, or install the latest medical technologies.

Asymmetric information and a failure of trust

Theoretically, nonprofit firms are not necessary to remedy shortfalls in government offerings or fulfill the goals of altruistic owners. For-profit firms could perform these functions as well; they could seek charitable donations and offer increased output while still making a profit. But Arrow (1963) argues that for-profit firms fail as charitable enterprises because of asymmetric information combined with a lack of trust. Donors may believe that for-profit firms are capable of high-quality charity work but do not trust these firms to use their donations exclusively for charity. It is difficult for donors to verify how exactly donations are put to use, and the firms have no reliable way to prove they are spending funds the way they claim. This lack of trust provides an opening for nonprofits to exist, because donors tend to trust the motivations of nonprofits more.

If this theory is correct, nonprofit firms should advertise their status whenever they interact with customers, because it is an advantage in attracting donations. However, David (2008) finds that few nonprofit firms in the hospital and nursing home sectors specifically call attention to their nonprofit status on their websites or in yellow page listings. This finding should be interpreted in light of the fact that donations make up only a small fraction of the operating budgets of firms in these industries.

Nonprofits as for-profits in disguise

All three of these theories assume that nonprofits and for-profits are different, and in particular that nonprofits offer more output either due to the altruism of their owners or their community. But are nonprofit and for-profit firms truly different, with different types of ownership, motivations, and ideals?

A fourth theory argues that nonprofit firms may just be for-profit firms in disguise (Brickley and Van Horn 2002). Recall that nonprofit firms are not permitted to have shareholders or generate positive monetary profits. This theory argues that owners and employees of the firm are effectively shareholders who capture the rents that would have been declared as profits by a for-profit firm. The "profits" are distributed not as stock dividends but rather as higher wages or non-monetary benefits like access to the company fleet. Under this theory, firms decide whether to organize as a for-profit or a nonprofit to maximize net returns to owners, shareholders, or employees based on the tradeoffs listed in Table 6.2 (**?**).

Lakdawalla and Philipson (1998) argue nonprofit status is actually a way for profit-maximizing firms to gain experience before switching to for-profit status. Fledgling firms initially organize as nonprofits and use the government tax breaks to lower operating costs. At this point, the inexperienced firm cannot generate substantial profits, so the costs of organizing as a nonprofit are small. Meanwhile, tax breaks enable them to expand output, which offers them experience and opportunities for learning by doing.

Once enough learning is accumulated and the firm can produce output more efficiently than its competitors, it may then switch to for-profit status to realize the newly possible profits. This hypothesis also explains why nonprofit and for-profit firms coexist in the hospital industry. The mixed production occurs because firms are at different stages of learning by doing.

Table 6.3. *Summary of theories for the existence of nonprofits.*

Theory	Argument
Government failure	Nonprofits exist to satisfy the demand for charity care above and beyond what the government provides
Altruistic motives	Some entrepreneurs have altruistic preferences (e.g. maximizing output, not profits) and organize nonprofits to achieve them
Asymmetric information	Nonprofits exist because donors cannot observe how for-profits will use their donations and do not trust them
For-profits in disguise	Nonprofits are actually profit-maximizing firms taking advantage of the legal benefits of nonprofit status

6.5 The relationship between hospitals and payers

Lastly, we discuss the economics of how hospitals are paid. In nationalized systems such as the UK and Sweden, the government decides how to fund hospitals. But in systems with private parties, payers typically negotiate with hospitals to decide the bill. We first introduce the idiosyncratic way that hospitals charge for their services in the US and then how those hospitals are actually paid. In Chapter 16 we discuss hospital payment in nationalized health systems in more detail.

The hospital bill

Hospitals bill for procedures according to a list of prices known as the charge description master or chargemaster. Each hospital manages its own chargemaster, and prices can vary tremendously from hospital to hospital, even within one region. In 2004, the chargemaster price for a chest X-ray ranged from $120 at San Francisco General to as much as $1,519 at Modesto Doctors Hospital (see Figure 6.3).

	SCRIPPS MEMORIAL LA JOLLA, San Diego	SUTTER GENERAL, Sacramento	UC DAVIS, Sacramento	SAN FRANCISCO GENERAL, San Francisco	DOCTORS, Modesto	CEDARS-SINAL, Los Angeles	WEST HILLS HOSPITAL, West Hills
Chest X-ray (two views, basic)	$120.90	$790	$451.50	$120	$1,519	$412.90	$396.77
Complete blood count	$47	$234	$166	$150	$547.30	$165.80	$172.42
Comprehensive metabolic panel	$196.60	$743	$451**	$97	$1,732.95	$576	$387.18
CT-scan, head/brain (without contrast)	$881.90	$2,807	$2,868	$950	$6.599	$4,037.61	$2,474.95
Percocet* (or Oxycadone hydrochloride and acetaminophen) one tablet, 5-325 mg	$11.44	$26.79	$15	$6.68	$35.50	$6.50	$27.86
Tylenol* (or acetaminophen) one tablet, 325 mg	$7.06	No charge	$1	$5.50	No charge	12 cents	$3.28

*Hospitals carry generic version, name brand, or both ** Represents the added total of 14 tests that make up the comprehensive metabolic panel
Sources: Scripps Memorial La Jolla; Sutter General; UC Davis Health Systems; San Francisco General; Doctors Medical Center; Cedars Sinal Health Systems; West Hills Hospital and Medical Center

Figure 6.3. *Prices for common procedures from California chargemasters in 2004.*

Source: Lagnado (2004). Reproduced with permission.

The list prices from these chargemasters, however, do not necessarily correspond to the amount that insurers or patients are actually charged. Private insurance firms as well as Medicare and Medicaid usually negotiate large discounts from hospitals, so the prices they pay are typically much lower than the chargemaster amounts. On the other hand, uninsured patients lack the market power to negotiate such discounts and can receive the full brunt of the chargemaster fees (Reinhardt 2006).

An article from *The Wall Street Journal* from 2004 relates the case of an uninsured Virginian patient who received a bill of $29,500 for a cardiac catheterization, a stent procedure, and a one-night stay in the hospital. In comparison, Medicare would have reimbursed $15,000 for the same procedure and Virginia's Medicaid only $6,000 (Lagnado 2004). This example of higher charges for an uninsured patient is not singular. In 2004 on average, the prices charged to the uninsured were about 150% more than the charges Medicare recognized, and this gap had been growing since 1984 (Anderson 2007).

Even in the case of the uninsured, what hospitals charge and what patients pay may still be different. Partially, this may be due to the inability of some uninsured to finance the full bill. But increasingly in recent years, spurred on by negative publicity, hospitals have been more proactive in providing discounts to the uninsured (Tompkins et al. 2006). For example, the Henry Ford Hospital System in Detroit, Michigan, charges uninsured patients $1,650 for a colonoscopy even though the estimated total cost is $2,750 (Henry Ford Health System 2012).

While different hospitals may charge different prices according to their own chargemaster, a single hospital can also charge different insurers different prices for the same procedure. The payment received by the hospital from a specific insurer for each procedure or group of procedures is typically determined through an annual negotiation between the hospital and insurer. The resulting price schedule can depend greatly on the bargaining position of hospitals, so the same hospital might bill different insurers very differently (Ginsburg 2010).

For example, "star" hospitals such as Massachusetts General Hospital in Boston and the Cleveland Clinic in Ohio command higher payments from insurers because of their stronger bargaining position. If insurers were to balk at these high bills and deny their customers coverage at these hospitals, many enrollees would switch to other insurers that do cover the star hospitals. For similar reasons, hospital systems that negotiate as blocs can also extract more from insurers, because they can control supply for a larger region. Evidence from US hospitals between 2002 and 2003 supports these hypotheses. Star hospitals earned $6,700 more per patient than non-star providers, and hospitals in systems generated $180,000 higher monthly profits than standalone hospitals. The latter finding may explain a recent movement toward hospital mergers (Ho 2009).

Uncompensated care

Chargemasters describe how much patients and their insurers are charged for medical care, but it does not always reflect how much hospitals are paid for procedures. For example, for some uninsured patients, even a discounted bill may be too costly to bear. If not paid, the care for these patients goes **uncompensated**.

Definition 6.5

Uncompensated care: hospital costs that are not covered by out-of-pocket payments, public insurance, or private insurance.

Studies indicate that as much as 6–7% of hospital expenditures are uncompensated. One study estimates that uncompensated care for the non-elderly in the US totaled $49.4 billion from 1996 to 1998, and Medicare estimates uncompensated care for the whole population reached $20.8 billion in 1999 (Hadley and Holahan 2003).

Who pays for uncompensated care? The answer is not "nobody," of course. Hospital bills that go unpaid must be either taken out of hospital profits, subsidized (explicitly or implicitly) by the government, or passed on in the form of higher prices to other customers. It seems likely that each of these players – Medicare, Medicaid, local and state governments, hospital shareholders, and other, non-indigent hospital customers – absorbs at least some of the costs of uncompensated care (Hadley and Holahan 2003).

In the US, both Medicaid and state governments provide explicit subsidies for uncompensated care. Medicaid disproportionate share (DSH) payments are given to hospitals that provide substantial uncompensated care. These payments can serve as an incentive for hospitals to increase their indigent care (Duggan 2000). Certain state governments also levy taxes on insurance payments to hospitals to fund a program dedicated to medical education and uncompensated care (Vladeck 2006).

Additionally, Medicare and some local governments make payments that are ostensibly for other purposes but effectively help defray the cost of uncompensated care. Like Medicaid, Medicare has a DSH program with payments allotted to hospitals serving low-income Medicare beneficiaries. Because Medicare beneficiaries, by definition, all have insurance coverage, this is not explicitly a subsidy for uncompensated care, but the same hospitals that tend to serve low-income Medicare beneficiaries also tend to accumulate the greatest number of unpaid bills.

Cost-shifting

More generally, even for care that is not uncompensated, hospitals may be engaging in *cost-shifting*, using richer recipients to subsidize poorer recipients of care. This hypothesis, also known as *cross-subsidization*, implies that patients with health insurance pay more for hospital care than the uninsured do. This practice is not at all novel. As early as 1958, health economist Reuben Kessel advanced the hypothesis that hospitals charge the rich more in order to finance the health costs of the poor (Kessel 1958). In the words of one surgeon:

> I don't feel that I am robbing the rich because I charge them more when I know they can well afford it; the sliding scale is just as democratic as the income tax. I operated today upon two people for the same surgical condition – one a widow whom I charged $50, the other a banker whom I charged $250. I let the widow set her own fee. I charged the banker an amount which he probably carries around in his wallet to entertain his business friends.
>
> (Seham 1956 p. 22)

Researchers have also tried to determine the extent of cost-shifting in hospital payments. Cost-shifting is hard to observe or measure because it is rarely explicit, so researchers focused on uncovering indirect evidence of cross-subsidization. For example, David et al. (2011) study the reaction of hospitals in Colorado and Arizona that face new local competition from cardiac specialty centers. Because cardiology and cardiovascular treatment is a major source of net revenue for the typical community hospital, increased competition from specialty centers is likely to reduce profits significantly. The study found that hospitals facing new competition reduced their provision of psychiatric services, substance abuse care, and trauma care. These are relatively unprofitable services that tend to be offered disproportionately to the uninsured population and often result in unpaid bills. This is suggestive evidence that hospitals use profits from lucrative services to subsidize uncompensated care.

If cross-subsidization is happening, which payers are subsidizing which others? Melnick and Fonkych (2008) study billing and payment data from California hospitals between 2001 and 2005. To compare the relative financial contributions of different payers, they calculate a *collection ratio*: revenue actually collected divided by the total chargemaster bill (their results for 2005 are listed in Table 6.4). The collection ratios for all groups – the privately insured, the uninsured, and those covered by Medicare or Medicaid – are far below

Table 6.4. *Collection ratio at California hospitals by insurance type, 2005.*

Insurance type	Collection ratio*
Private insurance	38%
No insurance	29%
Medicare	27%
Medi-Cal (California Medicaid)	23%

* The collection ratio is defined as the total revenue actually collected by the hospital divided by the official chargemaster bill for the procedures performed. These two numbers tend to differ due to negotiated discounts, patient bankruptcy, inability to pay, or refusal to pay.

Source: Data from Melnick and Fonkych (2008) analysis of Office of Statewide Health Planning and Development (OSHPD) Annual Hospital Financial Data.

100%, which means that hospitals collect far less for each procedure on average than the chargemaster would suggest. Additionally, the evidence that the collection ratio of those with private insurance is greater than among other groups suggests that patients with private insurance may be cross-subsidizing uninsured patients and patients with Medicare or Medicaid. Separate studies agree that cross-subsidization occurs in hospitals but argue that the effect tends to be small generally except in markets with less competition (Frakt 2011; Robinson 2011).

Meanwhile, in countries where the government pays for all medical care, such as the United Kingdom and Sweden, cost-shifting occurs well before the hospital visit. In these systems, hospital payments come not directly from patients at the point of service but instead from tax revenues. Progressive tax systems ensure that the burden of funding health care falls more on the wealthy than the poor.

6.6 Conclusion

Though hospitals may not have been the place to find effective cures at the start of the twentieth century, they are now technological marvels. Conditions that were once impossible to treat are now routinely cured, and surgeries that used to be unimaginable are now commonplace.

The technological advances of hospitals have not come cheaply: expenditures have exploded around the world. Some of the rise in costs reflects improvements in technology that are expensive but worthwhile, but a portion may be the result of hospital market power, the medical arms race, and the conflicting incentives of doctors and hospital administrators.

In efforts to constrain these costs, governments have adopted various policies that address the problems of market power, the positive and negative effects of market competition, and concerns about equity. We examine closely the tradeoffs different countries make in Chapters 15 through 17.

6.7 Exercises

Comprehension questions

Indicate whether the statement is true or false, and justify your answer. Be sure to cite evidence from the chapter and state any additional assumptions you may need.

1. Hospital admission rates have dropped in the past two decades.
2. There were more hospitals in the US in the late 1990s than there were in 1940; the largest source of growth has been among for-profit hospitals.
3. The average length of hospital stays in the US has remained flat after a sharp decline in the 1980s.
4. Higher values of the Herfindahl–Hirschman Index indicate higher levels of competition in those markets.
5. The hospital's experience with cardiac catheterization is at least as important as the cardiologist's experience in reducing complication rates following percutaneous coronary intervention (PCI).
6. Nonprofit firms can legally raise funds by issuing stock.

7 Consider the following theory due to Arrow (1963): nonprofits exist because for-profit firms are less trustworthy in the performance of actions that are hard to observe. According to this theory, government regulations requiring hospitals to report data on outcomes should lead to a lower share of nonprofit production in the hospital industry.

8 Medical arms races, since they are a form of private competition, lead to socially optimal levels of technology acquisition by hospitals.

9 Uncompensated care in the United States is almost entirely covered by government programs like Medicare and Medicaid.

10 Doctors are typically direct employees of hospitals in the US, whereas in the UK, they do most of their work in private practice settings.

11 In a DRG payment system, hospitals receive payment according to the number of services rendered.

Analytical problems

14 **Learning by doing.** Suppose a hospital is planning how many patients to treat over the next two years. One key factor the hospital needs to consider in its planning is that hospitals that see many patients in one year have lower costs of care in subsequent years. Let p be the price of hospital care, let q_1 be the number of patients the hospital treats in year 1, let q_2 be the number of patients the hospital treats in year 2, and let c_1 and c_2 be the hospital's per-patient cost in years 1 and 2. The hospital's profits over the two-year period are

$$\Pi = (p_1 - c_1)q_1 + (p_2 - c_2)q_2$$

The hospital takes p_1 and p_2 as fixed and they choose q_1 and q_2 to maximize profits, Π.

a Suppose, as a warm-up exercise, that there is no learning by doing, and that the unit cost functions are $c_1 = \frac{q_1}{2}$ and $c_2 = \frac{q_2}{2}$. Is the cost of caring for patients increasing or decreasing in the number of patients seen? Derive the supply function in periods 1 and 2 for the hospital. [*Hint:* Substitute the unit cost functions into the profit functions and take the derivative of Π with respect to q_1 and q_2.]

b Let us introduce learning by doing into the problem. Now, the unit costs in period 2 are lower if the hospital has a higher number of patients seen in period 1. As we have seen in this chapter, there is a lot of evidence that suggests that such learning by doing by hospitals does in fact happen. The unit cost function in period 1 is still $c_1 = \frac{q_1}{2}$, but the unit cost function in period 2 now depends also on the number of patients seen in period 1: $c_2 = \frac{q_2}{2q_1}$. Derive the supply function in periods 1 and 2 for the hospital. [*Hint:* Solve first for the optimal q_2 taking q_1 as given. Then use your question for q_2^* in your expression for q_1^*.]

c Should the hospital see more or fewer patients in period 1 if there is learning by doing?

15 **Herfindahl–Hirschman indexing.** The Herfindahl–Hirschman Index, H, gives economists a way of measuring how much competition there is in a market. Suppose

that there are N hospitals in a market, each with s_i percent of the market share, where $\sum_{i=1}^{N} s_i = 1$. Recall that the index is defined as follows:

$$H = \sum_{i=1}^{N} s_i^2$$

a Suppose that there are $N = 10$ firms in a market, and each of them has an equal share of the market, so that $s_i = \frac{1}{N} = \frac{1}{10}$ for each firm. What is the value of the Herfindahl–Hirschman Index in this market?

b Suppose instead that one firm in this market dominates such that it has 90% of the market share, while the remaining nine firms each have an equal share of the remainder. What is the value of the Herfindahl–Hirschman Index in this case? Do increases in H represent an increase or a decrease in competition?

c What is the largest value H can take? When is H equal to this number?

d Now suppose a new firm enters the market and pulls away market share from each of the incumbent firms such that $s_i = \frac{1}{11}$. Would you say that competition has increased with the entry of the new firm into the market? Calculate the value of H after the firm has entered.

e Under perfect competition, there are an infinity of firms in the market, each with an infinitesimal market share. What is the value of H in a perfectly competitive market?

16 Medical arms race. Consider a two-hospital town with hospitals unimaginatively named A and B. The hospitals are considering installing a new imaging machine to attract more patients. The population of the town is small enough that it really only needs one of these machines. Purchasing a machine costs $1,000.

- If both hospitals install a machine, each earns $800 in revenue.
- If only one hospital installs a machine, that hospital earns $1,800 in revenue. The other hospital loses $300 in revenue as patients switch to the other hospital.
- If neither hospital installs a machine, then revenue does not change for either one.

a Consider hospital A's decision to buy one of these fancy new machines. A's profits will depend on what B decides to do. Suppose hospital B buys a machine. How much will A earn if it also decides to buy a machine? How much will A earn if it does not? What should A do if B buys a machine?

b Now consider A's optimal choice if B decides not to buy a machine. How much will A earn if it buys a machine? How much will A earn if it does not? What is A's best response if B does not buy a machine?

c We assume that the hospitals make their decisions about the imaging machine concurrently, so this interaction can be modeled as a simultaneous-move game.

		Hospital B	
		Buy	Do not buy
Hospital A	Buy	A_{11}, B_{11}	A_{12}, B_{12}
	Do not buy	A_{21}, B_{21}	A_{22}, B_{22}

where A_{11} and B_{11} are hospital A's and B's respective payoffs when both hospitals install the imaging technology. Fill in this game payoff matrix.

d Hospitals A and B both act to maximize profits. What do you predict will happen in this market? Will the socially optimal outcome – only one machine is purchased in the town – happen? [*Note:* This type of game is commonly known as a Prisoner's Dilemma.]

e Suppose hospitals A and B share an owner who maximizes joint profits rather than individual ones. What do you predict will happen?

f In this case, is hospital competition good for welfare? What other information, if any, would you need to make an assessment?

Essay questions

17 Below is the abstract of a recent National Bureau of Economic Research working paper entitled "Can governments do it better? Merger mania and hospital outcomes in the English NHS" by Martin Gaynor, Mauro Laudicella, and Carol Propper (2011).

> The literature on mergers between private hospitals suggests that such mergers often produce little benefit. Despite this, the UK government has pursued an active policy of hospital mergers. These mergers are initiated by a regulator, acting on behalf of the public, and justified on the grounds that merger will improve outcomes. We examine whether this promise is met. We exploit the fact that between 1997 and 2006 in England around half the short term general hospitals were involved in a merger, but that politics means that selection for a merger may be random with respect to future performance. We examine the impact of mergers on a large set of outcomes including financial performance, productivity, waiting times and clinical quality and find little evidence that mergers achieved gains other than a reduction in activity. In addition, mergers reduce the scope for competition between hospitals.

a In the UK, most hospitals are owned by the government, rather than privately held. In a setting where most hospitals are not owned by the government (such as in the US), what effect do you predict that hospital mergers would have on the price of hospital care?

b Can you think of reasons why hospital mergers might lead to improvements in the quality of care for a given level of inputs (which is one measure of hospital productivity)?

c How might the fact that a hospital is run by the government in the UK change your prediction about the effect of hospital mergers on productivity?

18 Below is the abstract of a recent National Bureau of Economic Research working paper entitled "Human capital and organizational performance: evidence from the health care sector" by Ann Bartel, Ciaran Phibbs, Nancy Beaulieu, and Patricia Stone (2011):

> This paper contributes to the literature on the relationship between human capital and organizational performance. We use detailed longitudinal monthly data on nursing units in the Veterans Administration hospital system to identify how the human capital (general, hospital-specific and unit or team-specific) of the nursing team on the unit affects patients' outcomes. Since we use monthly, not annual, data, we are able to avoid the omitted variable bias and endogeneity bias that could result when annual data are used. Nurse staffing levels, general human capital, and unit-specific human capital have positive and significant

effects on patient outcomes while the use of contract nurses, who have less specific capital than regular staff nurses, negatively impacts patient outcomes. Policies that would increase the specific human capital of the nursing staff are found to be cost-effective.

General human capital includes the sorts of broad skills that people might learn during schooling; these sorts of skills apply to many kinds of work settings. By contrast, *specific* human capital includes the sorts of skills and knowledge that someone might learn while working on the job. Such skills cannot readily be translated to other jobs within the hospital, or even to the same job at another hospital. Why might contract nurses, who are typically hired for short periods of time by hospitals to fill unexpected scheduling holes, tend to have negative impacts on patient outcomes? Speculate about the sorts of specific human capital that might give regular nurses an advantage over contract nurses in helping patients. Can this help explain the positive volume–outcome relationship observed in hospitals?

Students can find answers to the comprehension questions and lecturers can access an Instructor Manual with guideline answers to the analytical problems and essay questions at **www.palgrave.com/economics/bht**.

This page intentionally left blank